

Kodikologie und Paläographie im digitalen Zeitalter 2

Codicology and Palaeography in the Digital Age 2

Schriften des Instituts für Dokumentologie und Editorik

herausgegeben von:

Bernhard Assmann	Alexander Czmiel
Oliver Duntze	Franz Fischer
Christiane Fritze	Malte Rehbein
Patrick Sahle	Torsten Schaßan
Philipp Steinkrüger	Georg Vogeler
Niels-Oliver Walkowski	Katharina Weber

Band 3

Schriften des Instituts für Dokumentologie und Editorik — Band 3

Kodikologie und Paläographie im digitalen Zeitalter 2

Codicology and Palaeography in the Digital Age 2

herausgegeben von | edited by

Franz Fischer, Christiane Fritze, Georg Vogeler

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Malte Rehbein, Patrick Sahle

2010

BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de/> abrufbar.

© 2011

Online-Fassung

Herstellung und Verlag der Druckfassung: Books on Demand GmbH, Norderstedt 2010

ISBN: 978-3-8423-5032-8

Einbandgestaltung: Johanna Puhl, basierend auf dem Entwurf von Katharina Weber

Satz: Stefanie Mayer und L^AT_EX

Inhaltsverzeichnis – Contents

Vorwort	VII
Preface	IX
Franz Fischer, Patrick Sahle Introduction. Into the Wide – Into the Deep: Manuscript Research in the Digital Age	XI

Digitale Reproduktion Digital Reproduction

Pádraig Ó Macháin Irish Script on Screen: the Growth and Development of a Manuscript Digitisation Project	3
Armand Tif Kunsthistorische Online-Kurzinventare illuminierten Codices in österreichi- schen Klosterbibliotheken	21
Alison Stones, Ken Sochats Towards a Comparative Approach to Manuscript Study on the Web: the Case of the <i>Lancelot-Grail</i> Romance	33
Melissa M. Terras Artefacts and Errors: Acknowledging Issues of Representation in the Digital Imaging of Ancient Texts	43

Digitaler Katalog und Semantik Digital Catalogue and Semantics

Silke Schöttle, Ulrike Mehringer Handschriften, Nachlässe, Inkunabeln & Co.: Die Erschließung der deutschen Handschriften und die Bereitstellung von Sonderbeständen in Online-Katalogen an der Universitätsbibliothek Tübingen mit TUSTEP	65
Marilena Maniaci, Paolo Eleuteri Das MaGI-Projekt: Elektronische Katalogisierung der griechischen Hand- schriften Italiens	75
Ezio Ornato La numérisation du patrimoine livresque médiéval : avancée décisive ou miroir aux alouettes ?	85
Toby Burrows Applying Semantic Web Technologies to Medieval Manuscript Research . . .	117
Robert Kummer Semantic Technologies for Manuscript Descriptions — Concepts and Visions .	133

Handschriften und Naturwissenschaften Manuscripts and the Sciences

Lior Wolf, Nachum Dershowitz, Liza Potikha, Tanya German, Roni Shweka, Yaacov Choueka Automatic Palaeographic Exploration of Genizah Manuscripts	157
Daniel Deckers, Leif Glaser Zum Einsatz von Synchrotronstrahlung bei der Wiedergewinnung gelöschter Texte in Palimpsesten mittels Röntgenfluoreszenz	181
Timothy Stinson Counting Sheep: Potential Applications of DNA Analysis to the Study of Medieval Parchment Production	191

Peter Meinlschmidt, Carmen Kämmerer, Volker Märgner Thermographie – ein neuartiges Verfahren zur exakten Abnahme, Identifizierung und digitalen Archivierung von Wasserzeichen in mittelalterlichen und frühneuzeitlichen Papierhandschriften, -zeichnungen und -drucken	209
---	-----

Digitale Paläographie Digital Palaeography

Peter A. Stokes Teaching Manuscripts in the Digital Age	229
Dominique Stutzmann Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ?	247
Stephen Quirke Agendas for Digital Palaeography in an Archaeological Context: Egypt 1800 BC	279
Markus Diem, Robert Sablatnig, Melanie Gau, Heinz Miklas Recognizing Degraded Handwritten Characters	295
Julia M. Craig-McFeely Finding What You Need, and Knowing What You Can Find: Digital Tools for Palaeographers in Musicology and Beyond	307

Transkription und Textkodierung Transcription and Text Encoding

Isabelle Schürch, Martin Rüesch Ad fontes – mit E-Learning zu ersten Editionserfahrungen	343
Carole Dornier, Pierre-Yves Buard L'édition électronique de cahiers de travail : l'exemple de Mes Pensées de Montesquieu	361
Samantha Saïdi, Jean-François Bert, Philippe Artières Archives d'un lecteur philosophe. Le traitement numérique des notes de lecture de Michel Foucault	375

Elena Pierazzo, Peter A. Stokes
Putting the Text back into Context: A Codicological Approach to Manuscript
Transcription 397

Appendices

Kurzbiographien – Biographical Notes 433
KPDZ 1 – CPDA 1 443

Vorwort

Das Institut für Dokumentologie und Editorik (IDE) legt hiermit den zweiten Band zur *Kodikologie und Paläographie im digitalen Zeitalter* vor, der zugleich als dritter Band in der institutseigenen Schriftenreihe erscheint. Seine Vorgeschichte ist kurz. Vor zwei Jahren, im Herbst 2008, startete das IDE eine Initiative, den Stand der Forschungen zum Einsatz von modernen Informationstechnologien in der Arbeit mit Handschriften zu dokumentieren. Im Juli 2009 konnte der erste Sammelband zum Thema auf einer internationalen Fachtagung in München präsentiert werden. Die Resonanz auf Band und Tagung fiel seither weitgehend positiv aus, gerade auch von den Experten aus der Handschriftenforschung, die noch wenig mit digitalen Arbeitsweisen vertraut waren. Denn zum ersten Mal ließen sich die eher disparat vorangetriebenen Ansätze und Ergebnisse der computergestützten kodikologischen und paläographischen Forschung als Gesamtphänomen ins Auge fassen und – auch kritisch – diskutieren. Gleichwohl lag es in der Natur der Sache, dass der damals vorgelegte erste Band zwar einen breitgefächerten Einblick in den State of the art geben, viele relevante Themengebiete und Fragestellungen jedoch nicht behandeln konnte. Das IDE entschloss sich daher, einen weiteren Band zu *Kodikologie und Paläographie im digitalen Zeitalter* folgen zu lassen. Dafür wurden einschlägige Experten gezielt angesprochen. Parallel dazu gingen über einen offenen Call for papers viele weitere Beitragsskizzen ein. In einem doppelten Peer-Review-Verfahren wurden die Vorschläge ausgewählt, die sich besonders gut in die Gesamtstruktur des Bandes einfügten.

Wir freuen uns nun, mit diesem zweiten Band die durch den ersten Band aufgeworfenen Fragen der Digitalisierung und Katalogisierung sowie der automatischen Schrifterkennung und Schriftanalyse vertiefen zu können. Das Themenspektrum ist inhaltlich durch Beiträge zum Computereinsatz bei der Erforschung von Musikhandschriften und Buchkunst und zur Anwendung naturwissenschaftlicher und neuer informationstechnischer Methoden erweitert worden. Auch konnte der Horizont über die Grenzen der lateinischen Schrift hinaus auf griechisches, glagolitisches, judeo-arabisches und anderes Schriftgut ausgedehnt werden. Der raum-zeitliche Rahmen spannt sich nun vom Ägypten des zweiten vorchristlichen Jahrtausends bis ins Paris der Postmoderne.

Verweise zu Webseiten und Online-Ressourcen in den bibliographischen Anhängen schliessen nach Möglichkeit Angaben zu Publikationsort und -zeitraum mit ein. URL-Adressen wurden generell Ende Oktober 2010 überprüft.

Unser Dank gebührt allen beitragenden Autorinnen und Autoren dafür, dass sie diesen Band überhaupt erst ermöglicht und sich in Anbetracht der knappen Fristensetzungen besonders kooperativ erwiesen haben. Darüber hinaus ist einer Reihe

von unentbehrlichen Helferinnen und Helfern herzlich zu danken: Randall Cream (Dublin), Katharina Mahler (Köln), Pádraic Moran (Galway), Eoghan Ó Raghallaigh (Dublin) und Aja Teehan (Maynooth) für inhaltliche Hinweise und muttersprachliche Korrekturen; Katharina Weber und Johanna Puhl (Köln) für die Einbandgestaltung und Bildbearbeitung; Torsten Schaßan (Wolfenbüttel), der an der Planung und ersten Umsetzung des Publikationsprojektes beteiligt war. Bernhard Assmann (Köln) bewältigte erneut alle technischen Feinheiten der Drucklegung. Die redaktionelle Mitarbeit unserer Kollegen Patrick Sahle (Köln) und Malte Rehbein (Würzburg) erstreckte sich auf alle wesentlichen Entwicklungsstufen dieses Bandes. Wir möchten uns ferner bei allen Archiven und Bibliotheken bedanken, die digitale Reproduktionen für die Ausgestaltung und Publikation bereitgestellt haben. Dank gebührt schließlich der Gerda-Henkel-Stiftung für die Unterstützung, die sie unserer Initiative hat zukommen lassen.

Dublin, Göttingen und München im November 2010, die Herausgeber

Anmerkung zur elektronischen Fassung

Diese Online-Fassung entspricht der gedruckten Fassung von 2010, die von Books on Demand vertrieben wird. Wir haben uns nur den Luxus geleistet, alle in Farbe vorhandenen Bilder auch in Farbe aufzunehmen. Der Satz ist leicht verändert, so daß der Seitenfall nicht immer genau der Druckfassung entspricht.

Dublin, Göttingen und Venedig im Juli 2011, die Herausgeber

Preface

The Institute for Documentology and Scholarly Editing (IDE) hereby presents the second volume on *Codicology and Palaeography in the Digital Age*, published as the third volume in the IDE series. The history is short. Two years ago, in autumn 2008, the IDE undertook to document the current state of computer-aided manuscript research. This initiative led to the launch of the first anthology on this subject at an international symposium in Munich in July 2009. The feedback on both the anthology and the symposium has been remarkably positive, from experts as well as from those not fully acquainted with digital methods. For the first time, widely dispersed, cutting-edge research in the field of digital codicology and palaeography could be surveyed as a whole, measured, and critically assessed. Yet, despite the fact that the first anthology gave a broad insight into theory and practice, many relevant topics and questions were not covered. For this reason the IDE decided to publish a second volume of *Codicology and Palaeography in the Digital Age*. Consequently, established scholars were asked for a contribution and an open call for papers was issued. All proposals received have been peer-reviewed.

To a certain extent this second volume deepens the questions raised by the first volume, particularly questions on digitisation and cataloguing, on character recognition and the analysis of script. At the same time the focus has been widened to now include the fields of computer-aided manuscript research in musicology and history of art as well as to methodologies applied in computational and natural sciences. The scope has been broadened from Latin writing to Greek, Glagolitic, Judeo-Arabic and other scripts. The spatio-temporal frame stretches from ancient Egypt of 1800 BC to Paris of the 20th century.

References to web sites and online resources in the bibliographies include, as far as possible, information on place and date of the publication. URLs were checked in late October 2010.

We are grateful to all contributors who made this volume possible and who under a very tight schedule proved highly collaborative. In addition, we received indispensable support at various stages in preparing this publication: We thank Randall Cream (Dublin), Katharina Mahler (Cologne), Pádraic Moran (Galway), Eoghan Ó Raghallaigh (Dublin) and Aja Teehan (Maynooth) for remarks and corrections on English texts; Katharina Weber and Johanna Puhl (Cologne) for image processing and cover design; Torsten Schaßan (Wolfenbüttel), who was actively involved in the initial planning and realisation of the project. Bernhard Assmann (Cologne) once again smoothly created the print version. Our collaborators Patrick Sahle (Cologne) and Malte Rehbein (Würzburg) had a significant influence at all stages of the editorial process. Moreover, we are

indebted to all archives and libraries who kindly provided digital reproductions for this publication. The IDE is grateful to the Gerda Henkel Foundation for the support granted to the *Codicology and Palaeography in the Digital Age* initiative.

Dublin, Göttingen and Munich, November 2010, the editors

Note on the Electronic Version

The online version of this volume is identical to the print version published by Books on Demand, Norderstedt 2010. However, greyscale images have been replaced with colour images where possible. Minor changes to the character font may have shifted pagebreaks slightly but general page references remain valid.

Dublin, Göttingen and Venice, July 2011, the editors

Introduction. Into the Wide – Into the Deep: Manuscript Research in the Digital Age

Franz Fischer, Patrick Sahle

Manuscript research is a wide field of scholarship which is integrated in core disciplines such as history, philology, or library science. Yet manuscript research is also crucial in other fields such as archaeology, history of arts, musicology or Egyptology, to name but a few. For all these disciplines, manuscripts are fundamental sources. There are different approaches to different types of manuscripts, but questions and perspectives, methodologies and tools are often quite similar. Innovations and new research strategies from one discipline can be transferred to and adopted by others. This introduction gives an overview of current aspects in the field of manuscript studies in both theory and practice by showing the relatedness of the contributions to the volume at hand as well as its predecessor, *Codicology and Palaeography in the Digital Age* (references given in parentheses). The texts are roughly assigned to five interrelated areas of manuscript research: (I) the photographic capturing of the manuscript surface, (II) the description of the manuscript for a catalogue, (III) the scientific examination of material aspects, (IV) the analysis of the script and (V) the deep encoding of the text itself.

I. Digital Reproduction

These days, the starting point for manuscript research projects is often digital reproduction. Digital facsimiles convey a great number of the original features and characteristics and can be easily provided and shared. Carried out on a large scale, digital reproduction is the cheapest way of making entire collections of manuscripts accessible. The criteria for selection vary according to research interests and institutions (cf. Kalning and Zimmermann in vol. 1).

The opening chapter of the first section, written by *Pádraig Ó Macháin*, is dedicated to one of the early digitisation projects, Irish Script on Screen (ISOS), initiated back in 1998 by the School of Celtic Studies at the Dublin Institute for Advanced Studies. This project's objective was, and still is, to digitise the entire Gaelic manuscript tradition—that is, all manuscripts in the Irish language—across all libraries and archives and to make the digital images freely available on the World Wide Web.

Research in art history can be greatly facilitated by means of manuscript facsimiles provided along with codicological data and descriptive texts. This can be more easily

achieved in the digital medium, as demonstrated by *Armand Tif*, whose chapter focuses on illuminated manuscripts from two particular monastic libraries in Austria. With the project described by *Alison Stones* and *Ken Sochats*, we leave the modern repository as the main organisational concept in favour of the manuscript tradition of one particular work, namely the popular Arthurian romance known as the Lancelot-Grail. Here, images and text are presented according to the narrative structure as well as the geographic dissemination of the manuscript witnesses. In a pre-analytical manner, this gives a promising starting point for comparative investigations beyond the flat surface of the reproduction.

An important caveat to manuscript research based on digital surrogates is articulated by *Melissa Terras*. Technical distortions can lead to the unintentional introduction of artefacts and errors into the digital representations of objects. These chapters show that digitisation is more than just a technical endeavour; it needs a methodology, and theoretical reflection upon the intersection of technical conditions and the requirements of critical scholarship.

II. Digital Catalogue and Semantics

On the other hand, any collection of digitised manuscripts would be of very limited use without inventories and catalogues indicating the content, material and provenance of each item in a particular collection. The new access to manuscript research by digital reproduction is still accompanied by a more traditional cataloguing approach. Codicology has always been the compilation and generation of knowledge about a manuscript, and cataloguing has been the dominant way of recording this knowledge. The creation of digital catalogues is increasingly common practice today, just as handwritten and print catalogues were common practice in archives and libraries before this (cf. Bernardi et al.; Cartelli et al.; Speer). Several software tools have emerged recently to facilitate this (for a description of just one such example, the M-Tool, see Uhlř and Knoll). An adapted version of the word processing program TUSTEP, described by *Silke Schöttle* and *Ulrike Mehringer*, proved to be an appropriate tool for the creation of the online catalogue for the special collections of the Tübingen University Library.

Converting knowledge from analogue to digital is not just a technical issue of how to do this as quickly and effectively as possible. Rather, it sets its own methodological agenda. One of the changes occurring in codicology, in comparison to traditional print cataloguing, is the relationship between the codex (or its visual digital surrogate) and its description (cf. Stinson) and therefore the description itself.

Moreover, there is a tendency towards open forms of collaboration in providing information and access to the manuscript heritage. The MaGI project presented by *Marilena Maniaci* and *Paolo Eleuteri* is an example of cataloguing facilitated by flexible digital tools and software (cf. Bernardi et al.; Cartelli et al.): across institutional boundaries, this project aims at cataloguing and selectively digitising all Greek codices held in Italian libraries.

The mere existence of digital reproductions and online catalogues prompts us to consider connecting catalogues by bringing all the available documentation together in comprehensive portals (cf. Uhlř and Knoll) and Virtual Research Environments (cf. Deckers et al.). Assessing the current state of digitisation, *Ezio Ornato* draws some radical conclusions. Based on the conviction that manuscripts are indeed written for the reader, his chapter reads like a codicological manifesto: researchers and cataloguers are called upon to unite as a community and to express their particular research needs, and databases must be created systematically and structured in order to realise the vision of a “*Bibliotheca universalis librorum Medii Aevi*”, freely accessible by means of a “*Catalogue grand ouvert*”. The condition for both of these would be a radical change and liberalisation in digitisation and publication policies of most of the manuscript libraries. This, in return, would have an impact on scholars who rely on a code of research ethics rigorously banning plagiarism, appropriation, forgery and obliteration.

Tendencies towards ever more comprehensive portals offer new opportunities for comparative studies and a global perspective on our cultural heritage. The next ‘evolutionary’ step in integrated manuscript descriptions is triggered by ideas from the ‘Semantic Web’. Here, the prevailing approach is to enrich already available data. This includes making semantically explicit what has been previously implicit in mere strings of characters. This means in turn that concrete objects as well as abstract concepts need to be identified in manuscript descriptions in order to connect these to entities from authority files and to bind them together via taxonomies and multilingual vocabularies. Yet the comprehensive usage of catalogue information across language borders and cultural practices in describing codices is only one side effect of ‘semantisation’. The chapters written by *Toby Burrows* and *Robert Kummer* both, independently from each other, sketch the basic concepts and current state of technical solutions for a “semantic codicology”. Both chapters reveal the enormous potential of semantic data and open a wide horizon for future research, where completely new questions may arise that scholars could not have previously imagined.

III. Manuscripts and the Sciences

In an ideal digital world, all knowledge of the handwritten tradition would be collected, connected, enriched and accessible from a single point of entrance. The sheer quantity

of all that easily accessible information might itself allow for qualitative progress in manuscript research. In addition, new approaches are emerging from the fields of information technology and the sciences. These too enable codicological research to gain new insights into the material aspects of cultural artefacts that have been already subjects of study for centuries.

Once manuscripts are available as digital facsimiles, we have the grounds for systematic analysis based on computational methods. Similarities and distinctions in scripts and individual hands can be measured and calculated (cf. Aussems and Brink). This sheds new light on the conditions and processes of manuscript production and on the number of scribes involved (cf. Stokes). Besides that, computational methods can also be applied to solve the problem of identifying fragments of documents that have been scattered and should be joined again. *Nachum Dershowitz, Yaacov Choueka, Roni Shweka* and *Lior Wolf* show how automated image analysis can produce significant new information and lead to well-founded suggestions about which fragments originate from a single document.

There is always more to a manuscript than meets the eye. An example of how hyperspectral imaging can be used to aid in text recovery is given in the first volume of this series (Shiel, Rehbein and Keating). The problem of faded and illegible writing is now addressed again in the chapter by *Daniel Deckers* and *Leif Glaser*. This contribution demonstrates how high-flux storage ring x-ray radiation can be applied to make script that has been erased visible again. Another ‘deep’ insight beyond multiple layers of written text and into the history of the supporting material can be gained by looking at the genetic makeup of the animal skin that is now parchment. The processed skin bears all the DNA information of the individual goat, sheep or calf from which it was taken. Systematic DNA sets of a large number of folia would fundamentally change the traditional way of dating and localising the creation of writing support, and the study of accidental characteristics would be completed by a scientifically grounded and possibly more reliable methodology. While the option of obtaining and analyzing such DNA strings has already been described in principle elsewhere, *Timothy Stinson* now ties the scientific approach back to the knowledge and evidence from the humanities again. Information on the DNA of parchments will contribute to codicological research and our understanding of parchment production as well as the history of animal husbandry. However, it always needs to be understood within the context of additional historical and archaeological evidence.

What the DNA is to parchment, the watermark is to paper. Watermarks give important indications regarding date and place of production of the writing support. *Peter Meinlschmidt, Carmen Kämmerer* and *Volker Märger* introduce thermography as a non-invasive method that yields clear pictures of watermarks. These images can be processed and integrated into comprehensive databases (cf. Wolf) in order to be compared and searched using pattern recognition techniques.

IV. Digital Palaeography

The growing mass of palaeographic information available online has changed the conditions for both research (cf. Ciula) and teaching (cf. Kamp) of what can now be called “digital palaeography”. Practical experiences in teaching are reflected upon in the chapter by *Peter Stokes*. How has the teaching of traditional skills changed, and to what extent should digital content be explicitly introduced into the curriculum for the study of medieval manuscripts? The author claims that a deep integration of both the digital and the traditional approaches has to be a fundamental principle, and that technical aspects are not a mere addition or something arbitrary. As such, palaeography should be taught in the wider context of Digital Humanities but at the same time “digital palaeography” should not be treated separately from palaeography in general.

In the tradition of a “quantitative palaeography” *Dominique Stutzmann*, by analysing medieval charters from Burgundy, demonstrates that the encoding of variant letter forms is an appropriate way to examine allographic characteristics (cf. Hofmeister et al.) and to draw conclusions about provenance and dating.

From the very beginning of palaeography as a discipline in the late 17th century, scholars have always been trying to classify scripts. This endeavour, again, has very much changed under the new terms and conditions of digital information and software tools (cf. Stansbury; Stokes; Aussems and Brink). Palaeography demands ever more detailed data and research (cf. Hofmeister et al.; Gurrado). The availability of information, and a transfer of methodologies, creates new possibilities for the study of handwriting as a cultural phenomenon across time and space. In recent years the interest in manuscripts and writing has increased beyond the occidental focus. This may comprise medieval oriental codices (contributions on this may be included in a planned volume III) as well as the study of texts from ancient Egypt under a palaeographic paradigm as introduced in the chapter by *Stephen Quirke*. Dating from about 1850–1750 BC, the several thousand fragments from Lahun form a promising collection of material for research that is based on computer-aided palaeography but that also aims at new insights into literacy and power in the ancient world.

Improving legibility is another major task in palaeography. Palaeographers, working mostly with digital surrogates of manuscripts, should make extensive use of advanced image processing techniques to improve the legibility of script for machine processing and human reading (cf. Fusi; Tomasi and Tomasi). *Markus Diem*, *Robert Sablatnig*, *Melanie Gau* and *Heinz Miklas* present their work on the pre-processing of images from Slavonic manuscripts in Glagolitic script, taking a new approach in applying OCR (optical character recognition) software to handwritten documents. Similarly, *Julia M. Craig-McFeely* presents techniques and tools for restoration and recognition of faded and degraded script in sources of special interest for musicologists. This contribution

goes even further, however, discussing digital tools for the next step in musicological research: the transcription of music notation and the creation of fluid scholarly editions.

V. Transcription and Text Encoding

In a similar manner, an attempt to bridge the gap between learning to read manuscripts (cf. Kamp; Cartelli and Palma) and learning to transcribe and edit digitally handwritten text is undertaken by the “Ad fontes” project as presented in the chapter by *Isabelle Schürch* and *Martin Rüesch*. But what is used here as a label for a project with pedagogical purposes in palaeography also describes a strong tendency in approaching our cultural heritage in general, and manuscripts in particular, under the new conditions of the digital age: *ad fontes*—to the sources! Manuscripts have become far more visible. And while the visibility and accessibility of manuscripts improve, their perception as documents in their own right—which are inseparably interconnected with their content—becomes more commonly accepted among scholars. As a matter of course, digital facsimiles are becoming an integral and natural part of scholarly editions. Two examples of these are the electronic edition of one of Montesquieu’s notebooks, presented in the chapter by *Carole Dornier* and *Pierre-Yves Buard*, and the digital archive of Foucault’s notes which he wrote when preparing “Les mots et les choses” (English: “The Order of Things”), as presented by *Philippe Artières*, *Jean-François Bert* and *Samantha Saïdi*. More than just illustrative examples, these projects also prove how documents can make visible the intellectual evolution of the thinking of these famous authors. With this in mind, and supported by a wide range of examples well worth looking at, *Elena Pierazzo* and *Peter A. Stokes* start to work out a revision of the prevalent perception of a manuscript text. In the guidelines of the Text Encoding Initiative (TEI), which may be regarded as the ‘de facto’ standard for the description and digital encoding of texts, the focus is primarily the text, a linguistic object, rather than the document, a physical object. In contradiction to this traditional attitude, the concluding chapter takes a codicological approach and argues for the establishment of a more document-centred markup standard for the transcription of manuscripts, that is, *putting the text back into context*.

Digitale Reproduktion



Digital Reproduction

Irish Script on Screen: the Growth and Development of a Manuscript Digitisation Project

Pádraig Ó Macháin

Abstract

Irish Script on Screen (ISOS), a project of the School of Celtic Studies at the Dublin Institute for Advanced Studies, was initiated in 1998, with the stated aim of the high-resolution digitisation of entire Gaelic manuscripts and of making the digital images freely available on the World Wide Web (www.isos.dias.ie). The growth and development of ISOS has therefore paralleled, and in some cases informed, the evolution of awareness of digital matters in Ireland over the last ten years. This paper describes the history and structure of ISOS, its public reception, its impact on research, and the varying uses that are made of the site. The questions of further potential and future direction are also addressed.

Zusammenfassung

Irish Script on Screen (ISOS), ein Projekt der School of Celtic Studies am Dublin Institut for Advanced Studies, war im Jahre 1998 mit dem Ziel initiiert worden, hochauflösende Digitalisate der gesamten irischsprachigen handschriften Überlieferung anzufertigen und diese im World Wide Web frei zugänglich zu machen. Das ISOS Projekt hatte dadurch in den vergangenen zehn Jahren maßgeblichen Anteil an einer erstarkenden computergestützten Forschung in Irland. Dieser Beitrag beschreibt Geschichte und Struktur von ISOS, öffentliche Wahrnehmung und Auswirkungen auf die Forschung sowie unterschiedliche Nutzungsformen der Website. Darüber hinaus werden Fragen zu dem Entwicklungspotential und den und Zukunftsperspektiven behandelt.

1. Origins and Objectives

Irish Script on Screen (ISOS) is a project of the School of Celtic Studies at the Dublin Institute for Advanced Studies (DIAS). It was initiated as a web-delivered digitisation project in 1998 in collaboration with the Department of Computer Applications (now the School of Computing) at Dublin City University (DCU), with DIAS as lead partner. This partnership came to an end in 2003, and the project is now run by DIAS alone.

DIAS is a statutory institute, and one of the primary functions of the School of Celtic Studies as laid down by law is ‘the investigation, editing, and publication of extant manuscript material in the Irish language’ (Institute for Advanced Studies Act 1940 §5.1a). Previous to the advent of digital technology, the School of Celtic Studies discharged this statutory obligation primarily through the publication of catalogues and editions, critical and diplomatic. The opportunities presented by the development of digital media in the 1990s for further prosecuting this duty were recognised at the time by Professor Pádraig de Brún (DIAS), an expert in Irish manuscript studies. If exploited properly and effectively, a new dimension could be brought to the study of Irish palaeography and codicology, a vision which anticipated very much the objectives and interests of the present publication. Irish philology in general would benefit in a dynamic way, as the whole package would be placed before an audience that could not have been imagined heretofore.

At a philosophical level, two principles were implicit in the project from the beginning. One was the right of everyone who wished to do so positively to have access to Ireland’s manuscript heritage. As DIAS is a public-service, supra-university research institution, it was agreed, in the spirit of enlightenment that informs the Institute’s operations and research, that access to the ISOS website and to its contents should be free and without unreasonable restriction; given the status of the Institute, and the nature of the subject matter, it was also agreed that the site should be bilingual (Irish and English), in so far as that was practicable. The second principle was one of partnership and collaboration. Collections of Irish manuscripts in libraries vary in extent and in antiquity: from small collections of nineteenth-century books representing products of the end of the scribal tradition, to large collections representative of both vellum and paper traditions. For ISOS to carry out its objectives, the project would enter into contractual collaborations with these libraries in a spirit of partnership, whereby the libraries would make materials specified by ISOS available for digitisation, in return for having those materials and their contents disseminated virtually through the website and for receiving copies of all images generated. From the outset, therefore, there was a feeling of shared purpose—reinforced by mutual generosity—in the creation of high-quality surrogates and in the sense of excitement that this important aspect of Ireland’s culture was suddenly being released from relative concealment.

The collaboration in the early years with DCU meant that, in addition to shared funding responsibilities, the work of the project was, at that time, divided between image-capture and description (DIAS), and processing, display and storage (DCU). In addition, it allowed the ISOS staff at the School of Celtic Studies to concentrate on the primary materials for the project, while technical questions of processing and storage were generally dealt with elsewhere. The good relations that were fostered between the collaborating partners, as well as the modalities of the day-to-day management of

the project, ensured that when the collaboration came to a natural conclusion in 2003, DIAS was well-positioned to assume all the functions of the project.

The objective of ISOS was, and remains, to digitise Irish manuscripts—that is manuscripts in the Irish language—cover to cover, and to make these digitised artefacts freely available to all through a dedicated website: www.isos.dias.ie. It was also intended, as a secondary consideration, that in time the digital images should be accompanied by ancillary material such as catalogues, commentary and transcripts, but that the capture and display of manuscript images would be the project's priority. It remains a basic principle that in a digitisation project which involves the creation of manuscript images, primary attention should be directed to the quality of those images. Websites can be upgraded again and again; digital text may be corrected, encoded and laid out in different ways over time; but the opportunity to digitise a given artefact may present itself only rarely.

2. Irish Manuscripts

Complete Irish manuscripts range in date from the early 12th century to the end of the 19th century. They are written in insular or Gaelic script—a remarkable continuity over eight centuries—and represent, in the texts that they contain, the full linguistic range from Old Irish to Modern Irish, including tracts on the *ogham* script that survived from the pre-literary period. Earlier manuscripts such as the Book of Armagh and the Stowe Missal (available on ISOS) contain prose material in the Irish language embedded in the otherwise latinate context of liturgical and ecclesiastical writings, while glossarial material and marginal verses in Irish occur in Latin manuscripts such as the St Gallen Priscian.¹ With some notable exceptions, manuscripts prior to 1600 were written on vellum; thereafter paper became the predominant material in Irish manuscripts. In all, it is estimated that about 5,000 Irish manuscripts survive today, the majority of them dating from the 17th century and later.

A word may be said at this point about the physical condition of these manuscripts, because the condition of the materials to be digitised has a bearing on the approach required for digitisation. In contrast to many other traditions, the amount of time spent in the comfortable custody of libraries by manuscripts of the Irish tradition is quite negligible. In other words, on account of their poor physical condition, these manuscripts are more in need of digitisation than most. Many manuscripts that were written in the vellum era, before the 17th century, were destroyed and lost in the wars and natural disasters of that period, and what survives represents a fraction of what once existed. The condition of many of these books betrays the poor treatment that they have been subjected to, by man and by the elements, throughout their history.

¹ Stiftsbibliothek, Cod. Sang. 904, available online at Codices Electronici Sangallenses (CESG).

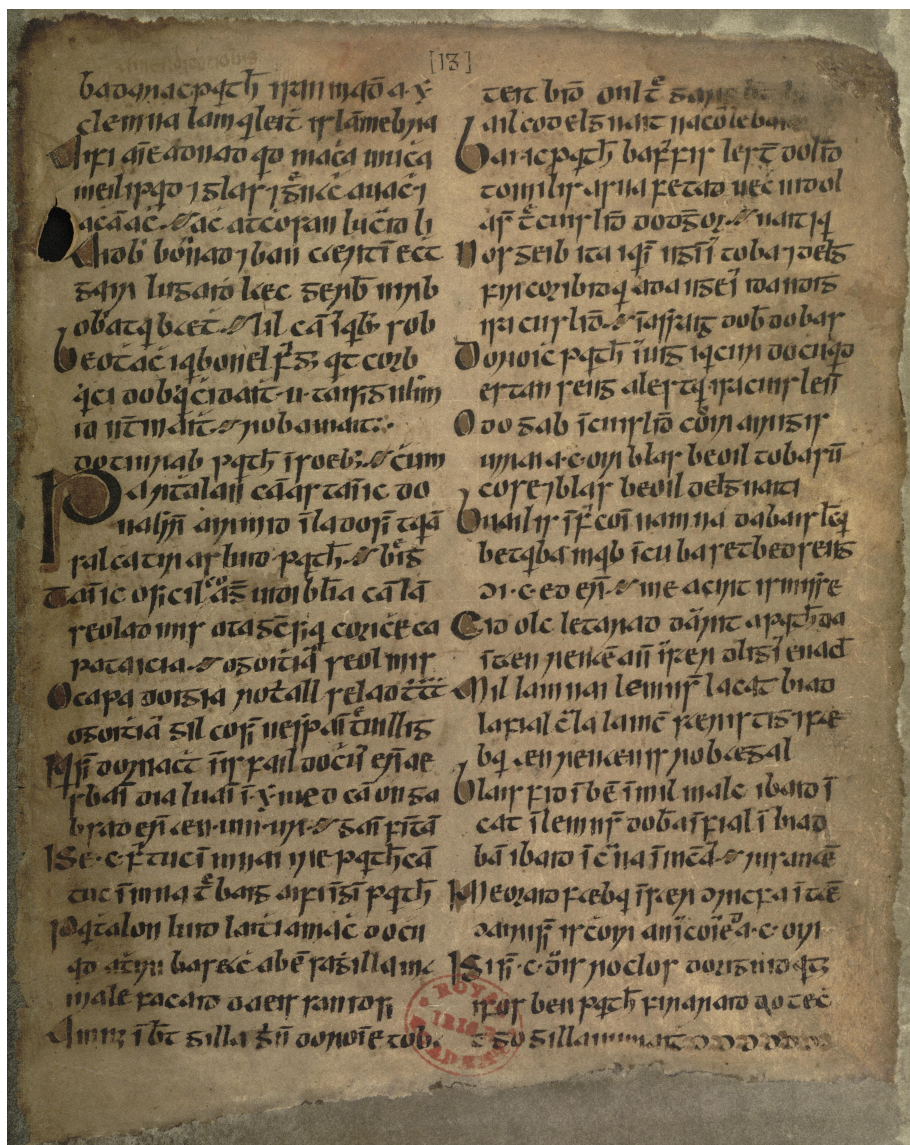


Figure 1. Page 13 of the Book of Fermoy (RIA 23 E 29), a fifteenth-century vellum manuscript, containing verse from *Leabhar Gabhála*.

The effects of damp, cockling, staining, and attacks by animals and man, are evident in many cases. As the codicological unit in medieval and late medieval Ireland was in many cases the *caidirne* (the quire or gathering), as opposed to the 'book', the practice of compiling manuscripts out of largely autonomous gatherings was widespread. The end folios of such gatherings often betray the fact that they were exposed, or even functioned as wrappers in addition to bearing text. Digitisation of vellum leaves, therefore, presents a challenge, particularly where the depth of field of the camera has to be fine-tuned to accommodate the uneven plain of the manuscript page: establishing a stable and uniform focus takes time and patience. By contrast, paper manuscripts are relatively easier to photograph, although, again, the condition in which such books were maintained in Irish tradition, prior to their reception in libraries in the 19th and 20th centuries, in addition to the varying properties of the inks used, means that the digitisation of such items presents its own challenges.

Prior to the advent of ISOS, access to Irish manuscripts was gained through the individual holding library, or through consultation of surrogates such as facsimiles (in a small number of cases only), transcripts, Photostat copies, or microfilm copies in black and white, and often of indifferent quality. When the ISOS project was initiated, the application of digital technology within the humanities was in its infancy. ISOS was the first web-delivered manuscript project in the digital humanities in Ireland, and, in the absence of any local precedent or template that might be followed, the project looked to the recently-established Celtic Manuscripts Project (now known as Early Manuscripts at Oxford University), which had been initiated in the Bodleian Library at the University of Oxford by David Cooper. In consultation with Cooper, the founder of ISOS, Pádraig de Brún, had begun the task of assembling the various components of the project prior to his departure from DIAS at the end of 1998, at which point planning and management of the project became the responsibility of the present writer. It is to the credit of de Brún that one of his many legacies to Irish scholarship will be the creation of this core research site in the field of Celtic Studies.

3. Protocols and Modalities

Taking account of the condition of the artefacts, and of the twin objectives of image-capture and web-delivery, it was decided that the optimum resolution for the digitisation of Irish manuscripts should be 600 dpi. This decision has proved over the years to be a sound one, and one that has stood the project in good stead. One consequence of this has been the consistency in image-quality from the beginning of the project to the present day.

Another decision taken during the test-stage of the project, and from which ISOS has benefited over the years, was that of two-tier delivery of the images to the users. It was

perceived that a balance needed to be struck between the requirements of the holding libraries and those of the users, and also between those of the general or casual user and the academic researcher. It was a salient and vital feature of ISOS from its inception that the project was user-driven: the principals who ran ISOS in DIAS were users of Irish manuscripts, and were thus able to envisage at first hand what would be needed and expected by those whom the website would ultimately serve. The primary requirement would be that of access to images that would, in the first instance, replicate as closely as possible the experience of viewing the manuscript page in its holding library. This type of access should be unhindered and, as far as the user's PC would allow, it should be immediate. Depending on the size of the manuscript page, a TIFF file—the file created in the first instance by the photographic process—might have a size of 130MB, which clearly would not be conducive to quick download and viewing by the end-user. It was also recognised that closer scrutiny of the manuscript text would be required, mainly by those with a scholarly interest and expertise in such matters. To cater for these two circumstances, a two-tier delivery system was devised, which is still in place.

Headers and footers are placed on the raw TIFF files, the header identifying the manuscript and the page or folio, the footer containing a copyright statement identifying the holding library. From these processed TIFFs, two JPEG files are created. The first is a 20% scaled version of the original image, with an average size of 200–400KB. This image can be downloaded quickly, and serves to represent the manuscript page as one might expect to see it in a library situation; it is accessible by everyone who visits the ISOS website, via a menu of GIF thumbnails. Feedback and experience show that, in practice, these small JPEGs are adequate for everyday use.

The upper tier consists of higher resolution images in the form of true JPEGs, the full-size, high-quality image which is designed for access by those engaged in scholarly research. These full-sized JPEGs can be in excess of 5MB. Access to these files is via the lower-tier images, with a username and password facility freely granted to applicants who download, complete and post by snail-mail a registration form available on the site. This is a mildly though deliberately inconvenient non-digital registration process, which implicitly demands that the user establish his or her bona fides by taking the trouble to register with the project in this way. All users are permitted to make a copy of the images for their own private use; they may also import the web-files into image-management software for the purpose of enhancing or otherwise manipulating an image, again for personal use only. ISOS does not involve itself in image enhancement, but tries to ensure that the image is a close representation of the appearance of the manuscript page in reality. It goes without saying that the magnification available through the high-resolution JPEGs, together with the potential for enhancement and manipulation through the agency of the end-user's image software, are what gives the digital image a life of its own.

There was another reason why this two-tier system was chosen over one of unrestricted access to an image of incrementally increasing magnitude. While most managers of repositories and most keepers of manuscripts are now accustomed to digital presentation as a desideratum and a fact of life, in the closing years of the 20th century such presentation was very much an untested innovation, particularly in Ireland. This gave rise to an understandable uneasiness among some library and archive professionals with regard to the exposure to a world-wide public of high-resolution images of hitherto concealed treasures. In those years, therefore, the two-tier system offered the re-assurance that access was not to be wholly unrestricted.

With the passage of time, the proliferation of digital delivery, and the promulgation of the benefits of mass digitisation, such apprehensions have waned to a great degree, and fear of the unknown is no longer a factor in decision-making. In addition, the obstacles of protectionism and elitism among custodians are nowadays encountered only very rarely. Nevertheless, the two-tier system has proved useful in allowing ISOS to form an impression of the public reception of the project, and of the types of use its images are put to among the scholarly community from whose ranks, virtually exclusively, the applicants for high-resolution usage are drawn. The communication between the project and the registered users also creates a conduit for valuable feedback.

Also surviving from the early days is another worthwhile feature of the project. This is the contractual basis which underlies each collaboration that is entered into. Again, in the early stages of the project, this held the extra value of re-assurance for potentially apprehensive curators. The true and lasting function of the contract, however, is the important and obvious one of setting out the modalities of the work (time-scale, manuscripts to be digitised), and the respective responsibilities of the partners with regard to practical matters such as insurance and copyright. In the case of the latter, copyright in the images is vested in the holding library, while that of ancillary material (texts, commentaries, catalogues) is vested in the author(s) of that material.

Finally, the important question of storage had to be addressed. It was decided that processed TIFF files should be stored on 40GB tape-cartridges, and that multiple copies of the tapes be created for storage in different locations. Because of the newness of the technology, however, manufacturers' claims regarding the capacity of these cartridges for long-life (30 years) storage were unverifiable, and at best were merely prognostic. It was decided that the tapes should be rewritten at three-year intervals, as a means of reviewing their ongoing well-being. In time, the project decided to move away from tape-storage (see below). The store of processed TIFF files constitutes the deep archive of the ISOS project.

4. Collaborations and Project-sets

Within a short time, collaborations were established with three of the primary repositories of Irish manuscripts in Ireland: Trinity College Dublin (TCD), the Royal Irish Academy (RIA), and the National Library of Ireland (NLI). Each of these projects was conducted in situ. This presented the perennial problem for all collaborating institutions: that of digitising-space. The ISOS work station requires an optimum space of 4 x 3 metres in order to accommodate camera, lights, book-cradle, computer and monitor, and to allow a measure of comfort to the digitising technician. Space is at a premium in all institutions, and only with difficulty have some institutions been able to allocate optimum space to the project.

The identification of target manuscripts is the responsibility of ISOS, after consultation with the holding library, and subject especially to the physical state of any given book. From the very beginning, certain project-sets suggested themselves, arising from areas of expertise within the School of Celtic Studies and from an understanding of the interests of the scholarly public. These project-sets influenced the selection of manuscripts in the early stages of the project. Three sets in particular were prioritised.

The first was the Great Books: the late-medieval and early-modern codices consisting of miscellanies of traditional learning and literature compiled from the 12th to the 15th centuries. Many, but not all, of these manuscripts carry identifying names, other than their library shelf-marks, names that have perpetuated their fame in the modern period. Among these are *Leabhar na hUidhre* (the earliest book written completely in Irish), the Book of Leinster, the *Leabhar Breac*, the Book of Lecan, and the largest format book that ISOS has digitised to date, the Book of Ballymote. Untitled manuscripts in this category include NLI MS G 2–3, one of the earliest manuscripts to survive from the post-Norman period.

The second project-set was the *duanaireadha*. These are manuscript anthologies of bardic court poetry composed by hereditary, professional poets in honour and in memory of members of prominent families in Ireland. The earliest such anthology to survive is the Book of Magauran (NLI MS G 1200), containing bardic poetry from the 13th and 14th centuries in honour of the family of Mág Shamhradháin of present-day Co. Cavan. One of the latest is a retrospective anthology of poems composed for the Ó Domhnaill family of Co. Donegal, compiled in the early 18th century (NLI MS G 167). Other *duanaireadha* which have been digitised include books containing poems addressed to the families of Ó Néill, Mac Suibhne, Ó hEadhra, Nugent, and De Róiste.

The third category was the scientific (mainly medical) manuscripts of the early-modern period. These books comprise a significant portion—roughly one quarter—of all extant vellum manuscripts in Irish, with some surviving also from the paper tradition. More than any other genre of Irish learned tradition, the contents of these manuscripts bear witness to the contact of Irish scholars of this period with European learning.

Most of the works are compilations and adaptations in translation of texts by the great authors of the time, such as Ibn Jazlah, Thaddeo Alderotti, and Bernard of Gordon. Of the three sub-categories established by the project, this was the one of which least was known among the general public, prior to digitisation.

While these project-sets were being undertaken, other partnerships were being formed among some of the constituent colleges of the National University of Ireland: University College Dublin (UCD), NUI Galway, and NUI Maynooth. Some very important books are housed in the college libraries, and it was important for the project to begin work in these repositories. Target manuscripts included a 17th-century medical manuscript in Galway (LSB 175), an early paper manuscript in the Russell Library at Maynooth (MS C 97), Mac Fhir Bhisigh's Book of Genealogies in UCD Library (Additional Irish MS 14), and the priceless Franciscan Collection of manuscripts held in the Department of Archives in UCD. Having identified primary target-manuscripts, the project was then in a position to work on other manuscripts from the collections in which those targets were to be found. In NUI Galway, for example, we were able to digitise the work of the important eighteenth-century scribes Labhrás Ó Fuartháin (LS 18) and Pádraig Ó Pronntaidh (LS 20). So too in Maynooth where, for instance, ISOS was enabled to digitise autograph manuscripts of poetry by Donnchadh Ruadh Mac Conmara (MS M 85) and Piaras Mac Gearailt (MS M 58(a)).

One of the natural results of the broadening of the collaborative base of the ISOS project was that further thematic categories began to emerge, some of which remain to be fully exploited. For example, digitisation of the recensions of the early-Irish saga *Táin Bó Cuailnge* in Leabhar na hUidhre and the Book of Leinster (RIA and TCD respectively) has led to the aspiration that manuscripts representative of all recensions of this text may yet be digitised and presented together on ISOS. A similar aspiration exists with regard to the manuscripts of the Annals of the Four Masters, of which the copies in the RIA and UCD have been digitised by ISOS. So too with manuscripts containing autograph copies of poetry. This is a particular feature of the later tradition, and, as indicated already, a number of books written by Irish poets and containing their own work survive from the eighteenth century, for example. In addition to the poets referred to above, the work of poets such as Aindrias Mac Cruitín and Seán Ó Murchadha is also represented in autograph manuscripts available on ISOS. It is envisaged that this sub-group will be augmented in the future.

The digitisation of the Franciscan manuscripts at UCD demonstrated further the capacity for the emergence of sub-projects within an open-ended, mass digitisation project. These manuscripts represent the core collection of the library of the Franciscan College of St Anthony in Leuven, Belgium.² Some of the manuscripts were written by

² A number of manuscripts from this source are also housed in the Bibliothèque Royale de Belgique in Brussels.

Irish exiles in Spanish Flanders and elsewhere, while more of the manuscripts were brought from Ireland by those exiles as they fled the country in the wake of the English conquest of Ulster at the beginning of the seventeenth century. Thus they range in date from the late 11th century to the 17th century and, in the case of later additions, the 18th century. The collection has had an interesting and precarious history: it was removed from Leuven to Rome in the wake of the French Revolution. From there the manuscripts made their way to Ireland in the nineteenth century, and were for many years housed at the Franciscan house of studies in Killiney, Co. Dublin. Their recent transfer to UCD and their digitisation by ISOS have brought their story full circle in a physical sense and also in a textual sense. Just as some of the manuscripts were studied at Leuven, and others created there, the possibility for further textual analysis and dissemination is now enhanced by their presence on ISOS. So too a sense of closure has been brought by this sub-project to the concern for preservation which caused the urgent movement of many of these books from Ireland and then across Europe.

Identifying target manuscripts among new collaborations and new sub-projects, in addition to those already established, took place in tandem with ISOS settling into a fruitful, long-term partnership with one of our early collaborators, the library of the Royal Irish Academy. Having digitised many of the great books in this collection, as well as the *duanairéadha* and medical manuscripts, it has become possible, due in no small part to the enlightened attitude of the Librarian, Siobhán Fitzpatrick, to begin working on a chronological basis with the Academy's manuscripts, the richest collection in existence. This work is in progress and it is envisaged that it will continue for some time.

The progress and growth of the project have recently encouraged us to enter into partnerships with institutions further afield. In 2009 the project concluded discussions with the National Library of Scotland. This library holds in excess of 70 Gaelic manuscripts, many of which originated in Ireland, and many more of which were written in Scotland. They date from the late middle ages to the modern period, and as a collection they have the extra significance of symbolising the closely shared heritage and cultural commerce of Ireland and Scotland. All the familiar categories previously prioritised by the project are present here: medical manuscripts, *duanairéadha*, and great books, perhaps the greatest of which is the Book of the Dean of Lismore, written in Scottish secretary script in Perthshire in the early sixteenth century. This manuscript is now on display on ISOS in conjunction with a new catalogue description by Ronald Black, who has spent a lifetime working on this collection.

Further still afield, ISOS has begun to pursue another sub-project, one which has the potential to develop in other directions in the future. This sub-project is neither text- nor genre-based, but rather seeks to emphasise the importance of the manuscript to the Irish emigrant. Wherever the Irish travelled—in the middle ages, the early modern period, or in the nineteenth century—they brought their books. In addition to the medieval

manuscripts of Irish origin or association scattered throughout Europe, we also find significant late manuscripts in holdings in the New World and in Australia. These books are testimony to the value placed by emigrants on the book as a cultural relic, and ISOS has begun to enter into collaborations with institutions in Australia so that some of these items may be digitised and displayed. Manuscripts from Newman College at the University of Melbourne and from the Benedictine Monastery of New Norcia are already on display.

The sense of purpose and mutual generosity that has informed these collaborations has led, naturally, to tangential work wherein, for example, ISOS has been pleased to facilitate the creation, on request, of digital copies of manuscripts not aligned with the core palaeographical concern of the project. In this way books such as the National Library's *Cambrensis* manuscript (MS 700), and the RIA's Icelandic medical manuscript (RIA MS 23 D 43) have found their way into the ISOS site. ISOS is also happy to act, when required, as a conduit for queries from publishers in other media regarding the use of ISOS images, and to supply those images once permission has been received from the copyright holders. Images generated by the project appear regularly in book and journal publications, in film documentaries, and as material in exhibitions.

5. Digital Developments: Technology

The speed of change over the last decade has affected ISOS in different ways. Filming times have improved, for instance. At present the project uses a single-shot digital back (Phase One P45+) with a large format 4 x 5 viewing camera (Sinar p2, with Sinaron lens). The first digital back used by the project, a Dicomed Studio Pro XL, worked as a scanner, which, depending on the size and nature of the manuscript page, might mean a significantly slower capture-time for a given image than would be the case today. Fortunately, capture-time does not affect image-quality, and the quality of the work that was done ten years ago compares very favourably with that of today's work.

Far and away the greatest advances, however, have been made in the area of digital storage, both in terms of cost and capacity. When the project began to generate data in 1999, a 10GB hard-drive was considered an object of wonder, and cost in the region of IR£145. Ten years later a 1TB hard-drive is commonplace and costs about €100. In 1999, if the project was working on a collection in situ, the raw images were conveyed from work stations to processing and storage stations on 1GB and 2GB JAZZ disks. These disks, and their drives, are now obsolete. The project has also witnessed and benefited from progress in the matter of bandwidth capacity over the years. In 1999 the capacity at DIAS was 64 Kbps. With the acquisition of a Digital Subscriber Line (DSL) this increased to 4Mbps, and the capacity is currently at 1Gbps via fibre-optic

cable. Users of ISOS have also seen an improvement in their own capacity to download and study images, as the now obsolescent dial-up connection gave way to DSL and 3G.

At present (January 2010), the ISOS archive consists of 4.5TB. Storage on 40GB tape-cartridges has not always proved satisfactory when reviewed and rewritten at the 3-year intervals which was established as a protocol when the project was at planning stage (see above). The archive has therefore been completely transferred to hard-drives, copies of which are contained on three separate servers, one of which is off-site. The ISOS website currently stands at 135GB. The working copy and the active site are both stored at different locations within DIAS, and a back-up copy is also housed off-site. When DIAS took over the exclusive management and operation of the project in 2003, two 2TB servers, costing €10,000 each, were purchased to store the project's archive. In 2009, hard-drives of 2TB capacity, using a fraction of the power and involving easier maintenance, cost €150 each. As developments in storage technology are taking place at a quicker pace than the generation of data within the ISOS project, it is likely that in a few years' time the complete ISOS archive may be accommodated on a single hard-drive.

6. Digital Developments: Attitudes and Perceptions

ISOS has grown in parallel with the development of digital awareness throughout the world, and particularly in Ireland. From a position where the project was innovative in conception, the ten years of its existence have witnessed an explosion in digital awareness and in digital thinking, and many libraries now have their own digital units. In many such cases the ISOS project has proved the guinea-pig for digital developments of this nature, and the project has often been requested to provide advice and guidance—which we have given freely—to other institutions and projects. This has happened in tandem with general technological—and behavioural—developments in areas such as the cell-phone and internet access. It has also been paralleled by the acceptance of digital technology as a conventional rather than an alternative medium for the dissemination of knowledge and information. In the context of Irish studies, this has been supported by the growth and development of important open-access public-service textual resources such as CELT (*Corpus of Electronic Texts*) and the electronic version of the Royal Irish Academy's *Dictionary of the Irish Language* (eDIL), and has led to the late flowering of many other similarly-minded enterprises such as the 1901 and 1911 Census of Irish Population, digitised from microfilm by the National Archives of Ireland, and the placenames database from the Placenames Commission.

In the context of professional scholarship, a digital component is now, tacitly or explicitly, the *sine qua non* of most government-funded third-level projects in the humanities. While this component, however, may be to the forefront in terms of

funding application and project presentation, in practice, in some cases at least, it has the status of a veneer. More than one project which received funding on the basis of a significant digital component, via the Irish government's Programme for Research in Third-Level Institutions (PRTL) or from the Irish Research Council for the Humanities and Social Sciences (IRCHSS), or indeed through a direct Government budgetary provision designated 'digitisation', has either prioritised delivery of its results in hard-copy format, or delayed the implementation of the digital commitment until the project-span and the funding have expired. Finding themselves in such a situation, some projects have sought help, on a *pro bono* basis, from public-service projects to execute belatedly the digital element of their work. The irony of such cases is that the funding for the nominally digital projects was granted after competition in which the public-service projects themselves were unsuccessful participants.

The perception of digitisation as peripheral, in certain quarters of the scholarly community, is one of the problems which has retarded the progress of the digital humanities in Ireland. Another area of concern is that of the allure of digital technology and digitised material for exhibition or 'front-page' purposes. In theory, this phenomenon should go hand in hand with and advance the cause of mass digitisation. In practice, the funds which might be directed towards facilitating such digitisation are instead diverted to sourcing turn-key and publicity-catching display technologies—touch-screen page-turners for instance—when a portion of this funding would have secured the digitisation of material the long-term value of which would exceed the short-term results accruing from the 'bells-and-whistles' display.

The issue is clearly one of balance. ISOS itself might be thought to be behind the times in its static display of images and catalogue material. Even though the current display-template for the project was established at a time when the progress and future of ISOS could not be predicted, nevertheless it still succeeds in achieving and delivering its objectives. Now that it has become settled and established, and that far more material continues to be added than was ever thought probable, it is not difficult to see that the potential shortcomings of static display are entirely rectifiable without prejudice to the digital content. Such re-organisation and upgrading tasks are among the desiderata for the project outlined at the conclusion of this paper.

The passing years have also witnessed the creation of what appears to be the parallel universe of European Union digitising theory and formulations, with their own set of acronyms and protocols. At a seminar organised by ISOS on the subject of 'Digital Image, Digital Text' in 2004, it came as a surprise to the personnel from the various projects in attendance to learn that Ireland had Nominated Representatives on the EU National Representatives Group for the Coordination of Digitisation Programmes and

Policies.³ The modicum of consultation and communication with the many digital projects in Ireland, which might be expected from such officials, has yet to materialise but will doubtless become apparent eventually.

It is regrettable that in attracting funding, short-term, all-inclusive, projects will undoubtedly be more successful than open-ended, infra-structural ones. Again, there is a balance to be struck between recognising the respective merits of projects that promise to deliver final results within a fixed period, and those that are laying down the basis for long-term research, the results of which, in so far as they can be predicted in any specific way, will be seen in areas—such as teaching, research and publication—over time rather than instantly. In the eyes of funding assessors, a digitisation project, the results of which are more likely to be external to itself, and to occur at an unspecified time in the future, is not going to outshine a self-contained short-term project with finite and verifiable deliverables.

7. Research

The research component of a project such as ISOS is generally potential rather than kinetic, and it increases as material is added. It is also accepted that time is required for the influence of a project such as ISOS to enter the public domain. Logfiles show that 27% of the visits to the ISOS site emanate from the .ie domain, and that there is also strong British, European and American interest in the project. One of the advantages of the registration mechanism for access to the high-resolution images is that we know, for example, that 48% of users who registered in 2009 were university postgraduate students, most of whom were working on studies of textual or palaeographical subjects; 37% of the registered users for the same twelve-month period were university staff. In addition, the availability of the mass of digitised material on ISOS has facilitated on-going research on binarisation, and on word and character segmentation, and the project is at present developing partnerships in this research area.

Some important research benefits are obvious within the site itself. Chief among the codicological benefits has been the virtual re-unification of manuscripts that became disturbed and fragmented over time, the separated parts now being housed in different institutions. For instance, Aoibheann Nic Dhonnchadha has identified the single leaf which is TCD MS 1398/71 as originally belonging to the sixteenth-century Maynooth MS C 110, and the two are now re-united on ISOS. The case of the Book of Leinster may also be cited. At least ten folios of this twelfth-century manuscript became separated from the codex at the end of the 16th century, and were conveyed to St Anthony's in Leuven, and hence ultimately to UCD, where they are now UCD-OFM MS A 3.

³ Since 2007 known as the Member States' Expert Group MINERVA, on which Ireland has four representatives.

The remainder of the book eventually became the property of TCD. These separated sections may now be viewed together on the ISOS site. The case of the Book of Leinster also illustrates the capacity for digitisation to prompt fresh scrutiny of a manuscript. An examination of the digital images from this codex led to the discovery of over 60 previously unnoticed marginalia (Manning 2003). Palaeography is also an obvious beneficiary from the ISOS images: John Carey's recent work (2009) on *Leabhar na hUidhre* (RIA MS 1229) is indebted to access to the digital images of that manuscript on the ISOS site, as is Pádraig A. Breatnach's forthcoming book on the work of the 'Four Masters'.

Many of the palaeographical and codicological advantages of digitisation are crystallized in a recent sub-project, the imaging part of which is already completed and on display. This involves the manuscript known as the Book of the O'Connor Don. This is an anthology of over 350 bardic poems that were composed in the period between the 12th and 17th centuries. The manuscript was written by an Irish soldier at Ostend in 1631. It extends to over 800 pages and is the most important and most comprehensive collection of this type of verse extant today.

In contrast to most other codices on display on ISOS, this manuscript is not housed in any national institution or educational library. It is one of the most important Irish manuscripts still to rest in private possession. Because of this, up to now access to the texts that it contains has been confined usually to a surrogate transcript made in 1848, or through a set of negative Photostat copies which were made in the first half of the 20th century. The owners always facilitated access, but few scholars ever journeyed to the heart of the Roscommon countryside to examine it in detail.

With the good will and agreement of the owners, ISOS was able to digitise the Book of the O'Connor Don in 2008, and to have the images on display by early 2009. Having a digital copy available of this manuscript has therefore immediately resolved and revolutionised the question of access to this important book. The most immediate result was that it enabled us, for the first time, to do such basic things as produce a catalogue of the contents of the manuscript: an alphabetical list of contents had been published in 1915. In the course of this work, other information became available for the very first time: previously unnoticed colophons by the scribe; other scribal features such as decoration, rubrication, and ruling; and extensive annotations and marginalia by later readers and owners of the 17th and 18th century.

As well as recording these new details and displaying them in conjunction with bibliographical information, work has begun on producing transcripts of each text and of displaying them side by side with the relevant images. This display is a feature additional to the usual two-tier display of the images. In this way a framework for diverse functions such as a palaeographical teaching-tool or a digital edition of some or all of the texts in the manuscript is being put in place gradually. Already a sample

digital edition of one of the texts is in preparation, as an ancillary feature of the ISOS site.

The Book of the O'Connor Don also illustrates the lateral properties of the ISOS site. The soldier-scribe of this manuscript, Aodh Ó Dochartaigh, is also the scribe of one of the Franciscan manuscripts (UCD-OFM A 20), a collection of less formal material telling the history of one of Ireland's greatest legendary figures, Fionn mac Cumhaill. This manuscript was written in 1626–7, also in Ostend, four years before the Book of the O'Connor Don. Now, for the first time scholars are enabled to study these two manuscripts side by side, from various aspects of codicology, palaeography, textual and literary history.

One of the more occasional features of the ISOS project has been the convening of seminars on various topics connected with digitisation. Over the years invited speakers have addressed topics such as presentation and storage of digital materials, text encoding, and digital editions. These seminars provide a forum for formal and informal interaction between those involved in digital studies in Ireland, and are also a means for ISOS to avoid the isolation that can beset many endeavours. Realising the importance of the digitisation of the Book of the O'Connor Don, ISOS convened a colloquium in 2009 on the subject of this manuscript, and of the numerous features and themes emerging from its sudden availability. This colloquium brought together many experts in the field of Irish manuscript studies and Irish classical poetry. The result was a completely new evaluation of this single source. It is intended to make the proceedings of this colloquium available in hard-copy and online in the very near future (Ó Macháin 2010).

8. The Future

In some respects, ISOS has become a victim of its own success and also of its venerable age. Neither the longevity of ISOS nor its enthusiastic reception and popularity could have been predicted when it was founded. As the oldest and consequently the longest-running project of its kind in Ireland it has seen the growth and the demise of better-endowed ventures, both within and without the academic environment. The project has made the most of every available opportunity to carry out its modest ambitions and to digitise manuscripts as they and the collections of which they form part became available. At any one time there have never been more than four personnel working on ISOS, and for many years now it has had just two persons: a digitising technician and the present writer.

For the future, the organisation of the website as a database is an obvious desideratum, and one which it is hoped can be attended to soon. Other aspects of the site are waiting for their potential to be realised. The static presentation of the material will, some time in the future, doubtless be exchanged for a more interactive facility. Another

requirement is the development of a schools' compartment in the site, which has been a personal ambition of this writer. A site such as ISOS is an ideal location to cultivate the next generation of Irish palaeographers by telling, without prejudicing scholarly standards, the story of the book, of vellum, parchment and paper, of ink and calligraphy, and of the descent of texts. As we look forward to such developments, we also anticipate new collaborations and sub-projects, and especially further additions to our digital collection of these unique and extraordinary hand-made books.

Bibliography

- Breatnach, Pádraig A. *The Four Masters and their works: a team enterprise*. Dublin: DIAS, forthcoming.
- Carey, John. "The LU text of *Lebor Gabála*". *Lebor Gabála Éireann: textual history and pseudohistory*. Ed. John Carey. London: Irish Texts Society, 2009. 21–32.
- Census of Ireland 1901/1911*. Dublin: The National Archives of Ireland, 2007–2010. <<http://www.census.nationalarchives.ie/>>.
- Codices Electronici Sangallenses (CESG) – Virtual Library*. Fribourg: University of Fribourg, 2005–2010. <<http://www.cesg.unifr.ch/>>. [The project is since 2008 part of e-codices, <<http://www.e-codices.unifr.ch/>>.]
- Corpus of Electronic Texts (CELT)*. *The Free Digital Humanities Resource for Irish history, literature and politics*. Cork: University College Cork, 1997–2010. <<http://www.ucc.ie/celt/>>.
- Early Manuscripts at Oxford University*. *Digital facsimiles of complete manuscripts, scanned directly from the originals*. Oxford: Oxford University and Oxford Digital Library (ODL) 2001. <<http://image.ox.ac.uk/>>.
- Electronic Dictionary of the Irish Language (eDIL)*. Royal Irish Academy (RIA) and University of Ulster, 2007–2010. <<http://www.dil.ie/>>.
- Institute for Advanced Studies Act*. Dublin: Rialtas na hÉireann / Oifig an tSoláthair 1940.
- Irish Script on Screen (ISOS)*. Dublin: DIAS, 1999–2010. <<http://www.isos.dias.ie/>>.
- Ó Macháin, Pádraig (ed.). *The Book of the O'Connor Don: essays on an Irish manuscript*. Dublin: DIAS, 2010.
- Manning, Gerald. "The later marginalia in the Book of Leinster". *Celtica* 24 (2003): 213–222.
- Ministerial Network for Valorising Activities in Digitisation, eContentPlus (MINERVA eC)*. 2006–2010. <<http://www.minervaeurope.org/>>.
- Placenames Database from the Placenames Commission*. Dublin: Rialtas na hÉireann / Fiontar (DCU), 2009–2010. <<http://www.logainm.ie/>>.

Kunsthistorische Online-Kurzinventare illuminierter Codices in österreichischen Klosterbibliotheken

Armand Tif

Zusammenfassung

Im Rahmen der Katalogisierungsprojekte illuminierter Handschriften und Inkunabeln am Otto Pächt-Archiv am Institut für Kunstgeschichte der Universität Wien wurden in den Jahren 2008–2010 zwei Web-Kurzinventare der mittelalterlichen Buchbestände in den Stiftsbibliotheken Herzogenburg und Stams erstellt. Der folgende Bericht stellt die dabei entwickelten kunsthistorischen Konzepte aus wissenschaftlicher und technischer Sicht vor. Der Schwerpunkt liegt auf den fachspezifischen Charakteristika der Kunstgeschichte in der Handschriften- und Inkunabelforschung. In technischer Hinsicht wird eine kostengünstige Lösung zur Web-Inventarisierung kleiner Sammlungen präsentiert.

Abstract

Two brief inventories of illuminated manuscripts and incunabula for online use were developed in the years 2008–2010 at the Otto Pächt-Archiv at the Art History Institute of the University of Vienna. The projects were concerned with the medieval book collections in the library of the monastery Herzogenburg in Lower Austria and in the library of the monastery Stams in Tirol. The following text presents the developed concepts focusing on the specific needs in the field of the history of art. A short technical description of a low budget solution for web-inventories of small collections is also included.

1. Einleitung

Gerhard Schmidt, der große Kenner österreichischer Handschriftenbestände, skizzierte 1996 eine Sondersituation in Österreich sehr treffend, wo – im Unterschied zu anderen europäischen Ländern – Klöster als Besitzer eines guten Teils mittelalterlicher Bibliotheken auftreten (Schmidt 7). Diese Besonderheit stellt auch die Handschriften- und Inkunabelkatalogisierung im digitalen Zeitalter vor Herausforderungen, die eigene Lösungen erfordern. Im Gegensatz zu den Sammlungen großer Institutionen wie Staats- oder Universitätsbibliotheken ist der Bestand eines einzelnen Klosters in der Regel zu

klein, um die Anschaffung und Verwaltung einer eigenen Datenbank zu begründen. Klöster brauchen daher professionelle Unterstützung, möchten aber ihrerseits – in ihrer Funktion als Besitzer ihrer Kulturbestände – als Partner wahrgenommen und der Öffentlichkeit entsprechend präsentiert werden (Penz).

Zwei Beispiele digitaler Kurzinventare von illuminierten Codices im Web, als Ergebnis gelungener Kooperationen zwischen Klosterbibliotheken und kunsthistorischer Forschung, wurden in den Jahren 2008/2009 (Augustiner-Chorherrenstift Herzogenburg in Niederösterreich, vgl. Tif/Roland) und 2009/2010 (Zisterzienserstift Stams in Tirol, vgl. Roland) im Kontext der Katalogisierungsprojekte des Otto Pächt-Archivs am Institut für Kunstgeschichte der Universität Wien erstellt. Die dabei entwickelten Konzepte sollen hier sowohl aus technischer als auch aus wissenschaftlicher Sicht vorgestellt und näher beleuchtet werden.

2. Web-Inventarisierung mittelalterlicher Buchbestände aus kunsthistorischer Sicht

In den letzten Jahren sind die Vorgaben des österreichischen Fonds zur Förderung der wissenschaftlichen Forschung (FWF)¹ hinsichtlich Digitalisierung und Web-Publishing in wissenschaftlichen Projekten zunehmend verstärkt worden. Der aktuellen Open Access Policy der finanzierenden Institution folgend, fiel 2008 im Rahmen eines Projekts unter der Leitung von Gerhard Schmidt der Beschluss, ein Web-Konzept für kunsthistorische Inventarisierung illuminierten Handschriften und Inkunabeln zu erarbeiten. Die grundlegende Überlegung war, nicht die Erfahrung aus der jahrzehntelangen Katalogisierungstradition der Buchmalerei in Wien (ausführlich: Roland 2009a) an eine vorgegebene Web-Präsentationsform anzupassen, sondern eine neue Web-Präsentationsform aus dieser Erfahrung heraus zu gestalten.

Reine Textkataloge bieten keine visuelle Suchfunktion für den unmittelbaren Bildvergleich. Durch fehlerhafte oder unscharfe Begriffsanwendung kommt es zudem häufig zu fachlichen Missverständnissen. Es war daher von Anfang an klar, dass kunsthistorische Web-Kataloge Abbildungen bieten müssen. Bilddatenbanken mit Einzelabbildungen vermitteln ihrerseits einen entkontextualisierten Eindruck des Buchschmucks, der für die aktuellen wissenschaftlichen Herangehensweisen der Handschriften- und Inkunabelforschung nicht genügen kann.² Die historischen Zusatzinformationen, die maßgeblich zur Lokalisierung und Datierung der Buchmalerei herangezogen

¹ Der Wissenschaftsfonds (FWF) ist die zentrale Einrichtung zur Förderung der Grundlagenforschung in Österreich und entspricht in etwa der Deutschen Forschungsgemeinschaft (DFG) in Deutschland. Die Katalogisierungsprojekte wie auch die Mehrheit der wissenschaftlichen Mitarbeiter am Otto Pächt-Archiv in Wien werden überwiegend vom FWF finanziert.

² Die wohl bekanntesten Beispiele im deutschen Sprachraum stellen die Bildarchive Prometheus und Bildindex der Kunst und Architektur dar.

werden, fehlen bei der Präsentation einzelner illuminierter Seiten oder gar nur einzelner Detailaufnahmen von Initialen beziehungsweise Randdekor. Der Kontext des Buchschmucks muss aber unbedingt nachvollziehbar sein.

In der kunsthistorischen Handschriften- und Inkunabelforschung kann die Arbeit am Original nicht durch Digitalisate ersetzt werden. Dennoch hat eine Vorauswahl des Materials durch bebilderte Online-Inventare einen konservatorischen Aspekt, da die Vorsortierung der zu bearbeitenden Codices nicht mehr an den Originalen erfolgen muss. Für die Langzeitarchivierung sind Volldigitalisate mittelalterlicher Codices unerlässlich. Dies können aber nur große finanzkräftige Institutionen durchführen. Die vollständige Digitalisierung ist für kleine und mittlere Sondersammlungen in Klosterbibliotheken nicht realisierbar. Auf der anderen Seite erzeugen Volldigitalisate auch große Datenmengen, die nur umständlich nach kunsthistorisch relevanter Information durchsuchbar und für den Forschenden nur mit erheblichem Zeitaufwand zu benutzen sind.

Die Entwicklung des Web-Konzepts zur Kurzinventarisierung sollte daher die spezifischen Kriterien zur Aufnahme kunsthistorisch relevanter Daten bei Objekt-Autopsien des Otto Pächt-Archivs berücksichtigen. Aus diesem Grund wurde entschieden, einen kleinen (Stams) und einen mittelgroßen (Herzogenburg) mittelalterlichen Stiftsbibliotheksbestand als Grundlage für die Erstellung eines Inventarisierungsschemas in digitaler Form zu verwenden. Beide Bestände wurden im Sommer 2008 durchgesehen und die mit Buchmalerei versehenen Codices kunsthistorisch autopsiert, wobei alle wichtigen Informationen in Digitalaufnahmen festzuhalten waren.³ In einem zweiten Schritt wurden die Metadaten erarbeitet. Die Erschließung von Texten, Schreiben, Einbänden, Provenienzen, Makulaturfragmenten, Buchschmuck, Sekundärliteratur etc. ist exemplarspezifisch vorgenommen worden. Erst danach sollten die Autopsieberichte als Textbeschreibung mit jeweils einer Bildergalerie zu jedem Codex erstellt und online veröffentlicht werden.

Die Usability der Web-Kurzinventare im Pächt-Archiv orientiert sich an den Bedürfnissen der kunsthistorischen Forschung. In der Funktionalität muss die Möglichkeit gegeben sein, mehrere Bilder zum Vergleich nebeneinander zu stellen. Des Weiteren sind persistente URL-Adressen sowohl für jede Bildergalerie als auch für jedes einzelne Bild einzurichten, um eine Direktverlinkung für wissenschaftliche Web-Publikationen zu ermöglichen. Um zu große Datenmengen zu vermeiden und dennoch sowohl einen Eindruck vom Gesamtlayout einer Manuskriptseite, als auch eine detaillierte Betrachtung des Buchschmucks zu gewährleisten, ist die Entscheidung getroffen worden,

³ Die Datenaufnahmen in der Stiftsbibliothek Stams wurden von Maria Theisen und Martin Roland durchgeführt. Für die Arbeitsschritte in der Stiftsbibliothek Herzogenburg zeichnen Martin Roland und der Verfasser verantwortlich. Zur Erstellung der Web-Kurzinventare wurden die digitalen Aufnahmen verwendet, die während der Arbeitsreisen in den zwei Klosterbibliotheken als Arbeitsmaterial entstanden.

mindestens je zwei Abbildungen (Gesamtansicht und Detail) von jeder Seite in die Bildergalerien zu stellen.

Es erschien als sinnvoll, die Zweiteilung in Text- und Bildband der in den letzten Jahrzehnten gedruckten Kataloge illuminierter Handschriften und Inkunabeln in Österreich auch für die Präsentationsform im Internet beizubehalten. Der wichtigste Vorteil liegt darin, dass der Text und die zugehörigen Abbildungen nebeneinander studiert werden können. Einen weiteren Vorteil bietet die Zweiteilung bezüglich der Durchsuchbarkeit des inventarisierten Materials. Während der Textteil nach Schlagworten durchsucht werden kann, ist im Bildteil eine konzentrierte visuelle Suche möglich.

Im Gegensatz zu allgemeinen Handschriftenkatalogen fokussieren kunsthistorische Kataloge auf formale und historische Informationen, die zur Datierung, Lokalisierung und Zuschreibung der dekorativen Ausstattung von Codices führen. Bei der Erschließung des Buchschmucks ist die selektive Präsentation des Bildmaterials und der Zusatzdaten in etwa zu Einbänden, Provenienzen und Texten entscheidend. Die kunsthistorisch relevanten Stellen in einem Codex werden während der Autopsie am Original ausgewählt. Diese Auswahl erleichtert die weiteren Bearbeitungsschritte bei der Katalogisierung. Nach stilistischen Kriterien werden Werkgruppen gebildet, welche die formalen Charakteristika eines bestimmten Buchmalerateliers aufweisen. Wenn es sich dabei um eine noch unbekannte Werkstatt beziehungsweise einen noch unbekannten Buchkünstler handelt, wird aufgrund logischer Zusammenhänge zwischen Texten, Schreibern, Einbänden und Provenienzen der Codices einer Stilgruppe eine möglichst genaue Datierung und Lokalisierung des Buchmalerateliers vorgenommen. Für eine gründliche Werkanalyse ist es somit unerlässlich, die essenziellen Basisdaten der jeweiligen Codices geschlossen einzusehen. Diese Möglichkeit können reine Bildersammlungen beziehungsweise Datenbanken mit Einzelabbildungen nicht bieten. Daher erschien eine Präsentation von jeweils einer Bildergalerie pro Codex mit begleitender Kurzbeschreibung als zusätzliche Textdatei als die sinnvollste Lösung für die wissenschaftlichen Bedürfnisse der kunsthistorischen Handschriften- und Inkunabelforschung.

Wenn es sich um Werke eines bereits erschlossenen Künstlerateliers handelt, so können sie diesem stilistisch zugeschrieben werden. Von besonderer Bedeutung für den Aufbau einer überzeugenden Argumentationslinie ist die verwendete Terminologie der spezifisch kunsthistorischen Beschreibung. Als Anschauungsbeispiel kann hier auf die von Maria Theisen vorgenommenen Zuschreibungen und Händescheidungen einzelner beteiligter Buchmaler der so genannten Prager Wenzelswerkstätten verwiesen werden, deren Wirkung bei der künstlerischen Ausstattung einer *Moralia in Job* (Herzogenburg, Stiftsbibliothek, Cod. 94 Bd. 1 und 2) und eines Gebetsbuchs des Leitomischler Bischofs Johann von Bucca (Stams, Stiftsbibliothek, Cod. 12) als gesichert

gelten kann.⁴ Vollbeschreibungen von auf bestimmte Buchmalerateliers spezialisierte Kunsthistorikerinnen und Kunsthistoriker sind allerdings nicht primäre Aufgabe von Kurzinventaren, deren Ziel es sein sollte, lediglich das Grundlagenmaterial gut sortiert und aufgearbeitet den jeweiligen Spezialisten für eingehende Forschungen zur Verfügung zu stellen. Auch hier gilt, dass Zuschreibungen und vor allem Händescheidungen mehrerer Buchkünstler, die oft an der Dekoration ein und derselben Handschrift beteiligt waren, nur bei geschlossener Einsicht in das Bildmaterial und in die historische Basisinformation möglich ist.

Gerade in kleineren Sammlungen ist die Wahrscheinlichkeit groß, dass mittels Web-Inventarisierung unbekannte Kunstwerke von teilweise hohem Rang neu entdeckt werden können. In der Stiftsbibliothek Stams wurde zum Beispiel ein 1482 datiertes Stundenbuch für das Kurzinventar aufgenommen, dessen Buchschmuck in weiterer Folge von Lilian Armstrong als Werk eines venezianischen Buchmalers mit dem Notnamen »Pico-Master« für einen Tiroler Auftraggeber erkannt und diesem zugeschrieben werden konnte.⁵ Es ist mit Sicherheit davon auszugehen, dass mit zunehmender kunsthistorischer Erschließung mittelalterlicher Bibliotheken im Web immer mehr solche Beispiele folgen werden.

3. Formate und Standards

Im Entwicklungsmodus des kunsthistorischen Web-Konzepts wurde zunächst für den Herzogenburger und in weiterer Folge für den Stamser Bestand je eine statische *html*-Seitenstruktur angelegt. Der Web-Auftritt im Rahmen der Homepage des Otto Pächter-Archivs ist somit an keine dynamische Datenbank gebunden, sondern als statische Lösung verwirklicht worden.⁶ Diese möchte primär als kunsthistorisches Modell zur Inventarisierung mittelalterlicher Buchbestände verstanden werden. Ein von Anfang an angepeiltes Ziel war dennoch die An- beziehungsweise Einbindung der Daten in die

⁴ Die vor kurzem erschienene Publikation von Theisen weist die Autorin als Spezialistin für böhmische Buchmalerei und im Besonderen für die am Ende des 14. Jahrhunderts für Wenzel IV. von Böhmen tätigen Werkstätten aus. Ihre Beiträge in den Kurzbeschreibungen zu Herzogenburg, Stiftsbibliothek, Cod. 94 Bd. 1 und 2 sowie Stams, Stiftsbibliothek, Cod. 12 in den entsprechenden, hier vorgestellten Web-Kurzinventaren geben einen guten Einblick in den kunsthistorischen Beschreibungsmodus für Kataloge illuminierter Handschriften.

⁵ Der Codex 44 der Stamser Stiftsbibliothek wird von Lilian Armstrong als Stams-Kneussl-Hours bezeichnet und es handelt sich um das bislang einzige Stundenbuch mit Buchmalerei des venezianischen Pico-Masters, der vor allem Inkunabeln ausgestattet hat. Vergleiche hierzu den Beitrag in den Kurzbeschreibungen zu Stams, Stiftsbibliothek, Cod. 44 im entsprechenden Web-Kurzinventar sowie die Ausführungen von Roland (2011 81).

⁶ Dass ein Webauftritt mit statischen Seiten auch für größere Bibliotheken und Sammlungen gewählt und erfolgreich geführt werden kann, zeigt das Beispiel der Abteilung für Sondersammlungen der Universitätsbibliothek Salzburg, deren Internetpräsenz seit Jahren von der Leiterin Beatrix Koll im hohen Maße professionell erstellt und betreut wird.

österreichweite Handschriftendatenbank der Kommission für Schrift- und Buchwesen des Mittelalters der Österreichischen Akademie der Wissenschaften *manuscripta.at*.⁷ Die kunsthistorischen Web-Kurzinventare möchten daher nicht als allein stehende Insellösungen verstanden werden. Sie bieten auf der einen Seite einen individuellen Web-Auftritt für die jeweilige Klosterbibliothek, liefern aber die Daten für die alle österreichischen Klosterbibliotheken umfassende Datenbank auf der anderen Seite. Zudem bieten sie eine Schnellübersicht des Bildmaterials mittels Thumbnail-Galerien zu den inventarisierten Codices, wobei in *manuscripta.at* hauptsächlich Textinformation vorgesehen ist und die Bilddateien nur durch Textverlinkung verknüpft sind. Auch Kostengründe sprachen für die gewählte statische Lösung. Zum gegenwärtigen Zeitpunkt ist eine Umsetzung der erarbeiteten Konzepte in dynamische CMS-Strukturen (Content Management System) angedacht.⁸

Die Durchsuchbarkeit im Web ist bekanntlich in erster Linie textorientiert, weshalb für die Metadaten *pdf* und *html* als Formate gewählt wurden. Auch im Sinne einer erhöhten Accessibility erschienen diese Standards als günstige Lösung. Zwischen den Bildergalerien mit minimalen Inhaltsangaben im *html*-Format und den ausführlicheren Kurzbeschreibungen im *pdf*-Format wurden Verknüpfungen eingerichtet. Die Zerteilung des Text- und Bildmaterials auf der Präsentationsebene setzt das erwähnte Erfordernis, um Text und Bild nebeneinander studieren zu können. Dabei ist auch die Möglichkeit gegeben, den reinen Text der Kurzbeschreibungen als *pdf*-Datei auszudrucken und in Verbindung mit den Online-Abbildungen zu lesen. Die Zerteilung hat für die Datenhaltung in technischer Hinsicht keine Relevanz.

Für den Abbildungsteil wurden ausschließlich *jpg*-Dateien verwendet, um das Datenvolumen bei guter optischer Darstellung gering zu halten. Die Bildgröße für die Web-Präsentation ist zwischen 1000 und 1200 Pixel für die Längsseite bei einer graphischen Auflösung von 72 dpi gewählt worden. Hochauflösende digitale Aufnahmen in druckfähiger Qualität wurden den jeweiligen Stiftsbibliotheken übergeben und können dort bei den zuständigen Personen in Verbindung mit Publikationsrechten angefragt werden.

⁷ Die von Alois Haidinger angesetzte Datenbank »Mittelalterliche Handschriften in österreichischen Bibliotheken« bietet eine Plattform auf nationaler Ebene für alle in Österreich aufbewahrten Handschriften und befindet sich in kontinuierlichem Aufbau. Dem Entwickler und Administrator sei an dieser Stelle für die gute Zusammenarbeit bei der Erstellung der kunsthistorischen Kurzinventare für Herzogenburg und Stams herzlich gedankt.

⁸ Ausgehend von den gesammelten Erfahrungen bei der Erarbeitung der hier besprochen Web-Kurzinventare plant Martin Roland gemeinsam mit Alois Haidinger ein großer angelegtes Projekt, welches die spezifisch kunsthistorische Eingabe und Suche in der gesamtösterreichischen Handschriftendatenbank der Österreichischen Akademie der Wissenschaften als Schwerpunkt vorsieht. Eine erste Vorstellung dieses Vorhabens fand im Rahmen eines Workshops von »e-codices – Virtual manuscript Library of Switzerland« (24. und 25. Juni 2010) in Fribourg (CH) statt.

Das Kurzinventar der Stiftsbibliothek Herzogenburg ist 2008/2009 als erstes realisiert worden, weshalb die Ergebnisse bereits auch über den entsprechenden Fonds in *manuscripta.at* abgerufen werden können.⁹ Das hierbei entstandene Konzept wurde für die Inventarisierung des Bestandes an illuminierten Handschriften der Stiftsbibliothek Stams adaptiert und weiterentwickelt, wobei vor allem die Usability verbessert wurde. Aus diesem Grund soll im Folgenden der konzeptionelle und technische Aufbau des optimierten Kurzinventars Stams näher erläutert und die Unterschiede zum Kurzinventar Herzogenburg gegebenenfalls gezeigt werden.¹⁰

4. Konzeptuelle Gestaltung und technische Umsetzung

Die Indexseiten einer Webpräsenz entsprechen dem Titelblatt einer gedruckten Publikation. Funktional muss eine Indexseite also den schnellen Einstieg zum gewünschten Inhalt gewährleisten. Während beim Kurzinventar Herzogenburg lediglich eine Auswahl zwischen deutscher und englischer Version zu finden ist, wurde die Indexseite des weiterentwickelten Kurzinventars Stams mit einer Navigation ausgestattet, die einen direkten Sofortzugriff zu jedem Inhalt ermöglicht.

Der weitere Aufbau des Inventarisierungsschemas entspricht in etwa einem gedruckten kunsthistorischen Katalog. Die Startseite enthält das Geleitwort zur Bestandsaufnahme mit Danksagung an alle beteiligten Personen und Institutionen. In der Einleitung wird die Geschichte des Bestandes kurz skizziert und auf ältere Kataloge verwiesen. Die Signaturenliste bietet eine Schnellübersicht der illuminierten Codices mit Kurztitel und jeweils einem Bildbeispiel pro erfasste Signatur. Als kunsthistorische Register sind die Übersichtstabellen zu verstehen, die unterschiedliche Werkgruppen nach Entstehungszeit, Entstehungsort, Kategorien des Buchschmucks und ikonographischen Kriterien bilden. Ausführlichere Beschreibungen zu den illuminierten Codices sind in einer Kurzinventar-Textdatei im *pdf*-Format zusammengefasst. Das Kernstück der Kurzinventare bilden allerdings die Bildergalerien, welche zu jedem illuminierten Codex die kunsthistorisch relevanten Autopsieaufnahmen zeigen. Aus diesem Grund wurde der Zugriff auf die Bildergalerien durch mehrere unterschiedliche Navigationsmöglichkeiten eingerichtet, welche je nach Bedarf oder Usergewohnheit alternativ benutzt werden können.

Die Generierung von informationshaltigen kurzen URL-Adressen für jedes Objekt einer statischen *html*-Seite basiert auf der richtigen Struktur der Ordner- und Dateinamen von Beginn an. Die hierarchisch angelegte Verwaltungsstruktur der Daten ergibt

⁹ Ein ausführlicher deskriptiver Projektbericht zum technischen Aufbau und Funktionalität des Kurzinventars Herzogenburg findet sich bei Tif (2011).

¹⁰ Eine detaillierte technische Anleitung zur kostengünstigen Erstellung von Web-Inventaren findet sich als »Low-Budget-Konzept zur Online-Inventarisierung von Kleinsammlungen« unter der Rubrik Materialien auf der Homepage des Otto Pächt-Archivs am Institut für Kunstgeschichte der Universität Wien.

schließlich den jeweiligen Link-Pfad in der Web-Publikation. Nach den Autopsien der illuminierten Codices am Aufbewahrungsort wurden die erzeugten Digitalaufnahmen deshalb in je einen nach der jeweiligen Bibliothekssignatur benannten Ordner pro Codex überspielt. Dadurch ergab sich für jede Handschrift ein Verwaltungsstring wie z.B. `.../ki/stams/cod_01`. Bei der Benennung der Bilder war darauf zu achten, dass der Dateiname in Kurzform der traditionellen Zitierweise Aufbewahrungsort – Handschriftensignatur – Seite entspricht. Eine Bildunterschrift »Stams, Stiftsbibliothek, Cod. 1, fol. 1r, Detail 2« wurde beispielsweise in »Stams_01_001r-d2« als Dateiname des digitalen Bildes übersetzt. Somit kann jede Abbildung auch im Fall eines Einzeldateidownloads richtig zugeordnet werden, weil die wesentlichen Informationen im Dateinamen eingebettet sind.

Eine Kombination aus kostenfreier und kommerzieller Software ermöglichte es, eine Low-Budget-Lösung für die technische Erstellung der Websites zu finden. Die Größenoptimierung der Bilder für das Web sollte mit einer Bildbearbeitungssoftware erfolgen, die über eine Funktion zur Stapelbearbeitung verfügt. Für die Web-Kurzinventare im Otto Pächt-Archiv wurde das kostenlose Programm »Free Batch Photo Resizer« verwendet.¹¹ Für die Erstellung der Bildergalerien empfiehlt es sich, einen Web Gallery Creator zu benutzen. Die Galerien für die Kurzinventare Herzogenburg und Stams wurden mit der kostenlosen Version von CD2HTML 3.4.2 erzeugt, weil dieses Programm einen sauberen *html*-Quellcode auch bei verschachtelten Tabellen generiert.¹² Im Kurzinventar Stams ist für jeden Codex eine eigene *html*-Seite generiert worden, die anschließend, mit der Bibliothekssignatur (`cod_01.htm`, `cod_02.htm` etc.) benannt wurde.¹³ Auf diese Weise ergibt sich für jede Handschrift eine URL-Adresse, welche die Aufstellung am Aufbewahrungsort im virtuellen Raum simuliert.

Auf Basis der Autopsieberichte wurden die kunsthistorischen Kurzbeschreibungen unter Berücksichtigung älterer Kataloge und gegebenenfalls weiterer vorhandener Publikationen geschrieben. Als Textverarbeitungsprogramm ist MS-Office Word verwendet worden. Die gesamte Word-Datei konnte abschließend problemlos in ein *pdf*-Dokument umgewandelt werden, in der die gesetzten Links funktionsfähig bleiben. Neben dem

¹¹ Das Programm wird von RealWorld Graphics zum freien Download zur Verfügung gestellt. Auf der Homepage des Anbieters kann auch eine ausführliche Anleitung mit detaillierten Erläuterungen zu den Einstellungen und Funktionen eingesehen werden.

¹² CD2HTML 3.4.2 ist als Freeware über die Homepage des Herstellers Falk zu beziehen.

¹³ Bei der Erstellung des Kurzinventars Herzogenburg wurde eine Ankerpunkt-Lösung gewählt, um eine schnelle Übersicht des gesamten Bestandes zu erzeugen. Die Benutzung der Seiten im Web hat allerdings gezeigt, dass das große Datenvolumen der Thumbnails pro Seite teilweise zu einem langsamen Aufbau des Inventars führt. Des Weiteren wird die externe Verlinkung von anderen Websites auf die Ankerpunkte je nach verwendeten Browser unterschiedlich gelesen, weshalb es zu unterschiedlichen Darstellungsweisen der abgerufenen Ziele kommt. Um dies zu vermeiden, ist beim Kurzinventar Stams auf die Ankerpunktlösung verzichtet worden, auch wenn die Erstellung einer eigenen *html*-Seite pro Codex eine breitere *html*-Struktur ergibt.

ebenfalls im *pdf*-Format eingescannt, älteren Bestandskatalog der Stiftsbibliothek Stams und den *html*-Seiten zu Einleitung, Signaturenliste und Übersichtstabellen befinden sich die Kurzbeschreibungen als ausführlicher Inventartext auf der mittleren Verwaltungsebene des Kurzinventars. Minimale notwendige Informationen zu den illuminierten Codices sind den Beschreibungen entnommen und in die jeweiligen *html*-Seiten mit den Bildergalerien eingefügt worden.

Als letzter Schritt erfolgte die Einrichtung der Navigation mit einem *html*-Editor, wobei hier die professionelle Software Dreamweaver zum Einsatz kam. Das Programm wurde auch zur Erzeugung der Indexseite und der weiteren *html*-Seiten ohne Bildergalerien verwendet. Sowohl beim Kurzinventar Herzogenburg als auch beim Kurzinventar Stams wurden mehrere alternative Navigationsmöglichkeiten eingebaut, die sowohl das punktuelle Springen via *drop-down*-Menu als auch das »Blättern« wie in einem gedruckten Katalog erlauben.

5. Externe Verknüpfungen und Datenmigration

Der Zugriff auf die Indexseite kann sowohl über die Seite des Stiftes als Besitzinstitution als auch über die Seite des Otto Pächt-Archivs erfolgen. Selbstverständlich ist eine Verlinkung auf die Indexseite oder nur auf bestimmte Inhalte auch von jeder anderen Website möglich. Ein gelungenes Beispiel dafür zeigt die Kooperation mit *manuscripta.at*¹⁴, wo die Daten des Kurzinventars Herzogenburg bereits eingearbeitet wurden.

Da die meisten Verweise in den statischen *html*-Seiten relativ aufgebaut sind und nur eine einfache Codierung verwendet wird, können diese ohne Probleme in ein dynamisches CMS überführt werden. Sollte das CMS über einen integrierten *html*-Editor verfügen, ist die Nachbearbeitung der *html*-Dateien innerhalb der Datenbank möglich. Die Verlinkung der Bildergalerien in den *pdf*-Dateien weist absolute Links auf, welche allerdings funktionsfähig bleiben, sofern die URL-Adressen der Seiten nicht ohne die Einrichtung einer *redirect*-Direktive geändert werden.

6. Rezeption in der Fachwelt

Das kunsthistorische Web-Kurzinventar der Stiftsbibliothek Herzogenburg ging im April 2009 online. Eine offizielle Präsentation folgte im Juni desselben Jahres im Rahmen eines Festakts zur Neueröffnung der restaurierten barocken Bibliothek des Stiftes. Die Benutzungsstatistik für 2009 zeigt, welche Resonanz das Pilotprojekt des Otto Pächt-Archivs seitdem erfahren hat. Die hohen Besucherzahlen in den Monaten Juni, Juli und August sind natürlich auf die Wirkung der festlichen Präsentation im

¹⁴ Siehe Anm. 8.

Usage Statistics for paecht-archiv.univie.ac.at

Summary Period: Last 12 Months
Generated 19-Jan-2010 10:51 CET

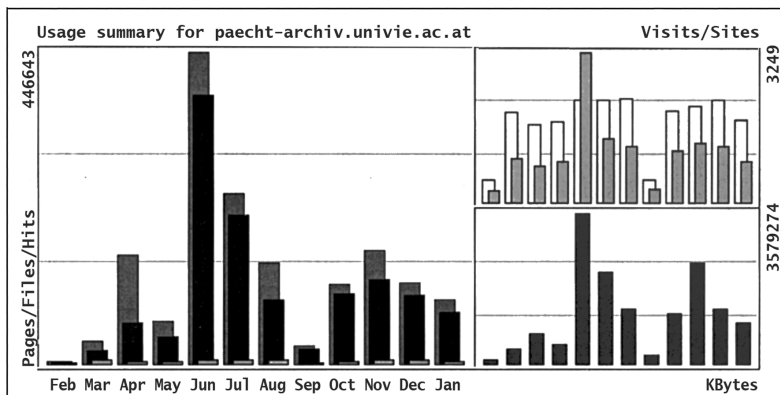


Abbildung 1. Usage Statistics for paecht-archiv.univie.ac.at.

Stift zurückzuführen, die in den österreichischen Medien angekündigt wurde. Der angezeigte Einbruch im September ist kein realer, sondern einer Fehlfunktion des Statistikprogramms geschuldet, das in jenem Monat einen wochenlangen Datenverlust aufweist. Von größerer Relevanz sind daher die statistischen Werte in Oktober, November, Dezember und dem darauf folgenden Januar 2010. Diese Werte blieben seither relativ konstant und zeugen von einer Vervielfachung der Besuche auf der Website des Pächt-Archivs. Das Kurzinventar Stams ist erst seit Mai 2010 online, weshalb hier noch keine entsprechenden statistischen Daten vorliegen.

Einer der Gründe für den schnellen Anstieg der Besucherzahl liegt mit Sicherheit in der Ausrichtung der Website auf die textorientierte Suche. Damit sind nicht nur die verwendeten Formate, sondern ist vor allem die fachspezifische Terminologie gemeint.¹⁵ Diese ermöglichte in kürzester Zeit eine hohe Platzierung der Ergebnisse über Websuchmaschinen wie zum Beispiel Google zu erzielen. Die Beschlagwortung der Inhalte mit wenigen präzisen Termini, unter Berücksichtigung technischer, formaler,

¹⁵ Die hier besprochenen kunsthistorischen Web-Kurzinventare konnten von der langjährigen Erfahrung von Martin Roland profitieren, der die Anwendung fachspezifischer Termini zur Beschreibung von Buchmalerei bei der Handschriftenkatalogisierung als einen seiner Interessenschwerpunkte untersucht. Stellvertretend für die Publikationstätigkeit auf diesem Gebiet sei hier verwiesen auf seine Mitarbeit an der Neuauflage des Standardwerks von Christine Jakobi-Mirwald.

hierarchischer und ikonographischer Aspekte des Buchschmucks, darf als einer der zentralen Punkte bei der Katalogisierung von illuminierten Codices betrachtet werden. Durch die Erstellung der kunsthistorischen Übersichtstabellen wurden die Kurzinventare mit einer digitalen Registerfunktion versehen, die das Fachpublikum zu einer hohen Trefferquote über die Textsuche im Web führt.

Die zahlreichen beschriebenen und bebilderten Beispiele zeigen aber auch den nicht kunsthistorisch ausgebildeten Kodikologinnen und Kodikologen, was als Buchschmuck definiert wird und wie die terminologisch richtige Bezeichnung gewählt werden sollte. Die Informationen aus den kunsthistorischen Kurzinventaren können daher auch für allgemeinere Kataloge von Nutzen sein. Zudem zeigt die Erfahrung, dass die Datierung und Lokalisierung der Buchmalerei oft wichtige Anhaltspunkte zur Datierung, Lokalisierung und Bestimmung von Schreibern, Texten, Einbänden etc. liefern können. Aus interdisziplinärer Sicht soll der Informationsaustausch über das Internet angestrebt und erleichtert werden, weshalb auch die Web-Kurzinventare eine Kontaktmöglichkeit mit den Bearbeiterinnen und Bearbeitern aufweisen, die für jede wissenschaftliche Zusatzinformation oder neue Erkenntnis zum inventarisierten Kulturgut dankbar sind.

Bibliographie

Deutsches Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg. *Bildindex der Kunst und Architektur*. [Marburg: Bildarchiv Foto Marburg, 2004–2010.]

<<http://www.bildindex.de/>>.

DFG: *Die Deutsche Forschungsgemeinschaft*. [Bonn: Deutsche Forschungsgemeinschaft, 2010.]

<<http://www.dfg.de/>>.

Falk, Petro. *CD2HTML*. [Cottbus, 2000–2003.] <<http://www.cd2html.de/>>.

FWF: *Der Wissenschaftsfonds in Österreich*. [Wien, 2010.] <<http://www.fwf.ac.at/>>.

Institut für Kunstgeschichte der Universität Wien. [Wien: Universität Wien, 2010.]

<<http://kunstgeschichte.univie.ac.at/>>.

Jakobi-Mirwald, Christine. *Buchmalerei. Terminologie in der Kunstgeschichte*. Berlin: Reimer, 2008.

Koll, Beatrix. *Universitätsbibliothek Salzburg, Abteilung für Sondersammlungen*. [Salzburg: Universitätsbibliothek, 2009.] <<http://www.ubs.sbg.ac.at/sosa/webseite/sosa.htm>>.

KSBM: *Kommission für Schrift- und Buchwesen des Mittelalters der Österreichischen Akademie der Wissenschaften*. [Wien: Österreichische Akademie der Wissenschaften, 2009.]

<<http://www.ksbm.oeaw.ac.at/>>.

manuscripta.at: *Mittelalterliche Handschriften in österreichischen Bibliotheken*. Kommission für Schrift- und Buchwesen des Mittelalters der Österreichischen Akademie der Wissenschaften. [Wien: Österreichische Akademie der Wissenschaften, 2009.]

<<http://www.manuscripta.at/>>.

Prometheus: Kunsthistorisches Institut der Universität zu Köln. *Bildarchiv Prometheus*. [Köln: 2001–2010.] <<http://www.prometheus-bildarchiv.de/>>.

- Otto Pächt-Archiv am Institut für Kunstgeschichte der Universität Wien. [Wien: Bibliotheksstiftung Otto Pächt, 2005–2010.] <<http://paecht-archiv.univie.ac.at/>>.
- Penz, Helga. *Klosterbibliotheken und Forschung: auf dem Weg zur Partnerschaft*. 14th International Congress "Cultural Heritage and New Technologies", 2009 66–67. <http://www.stadtarchaeologie.at/wp-content/uploads/eBook_WS14_Part2_Workshops.pdf>.
- RealWorld Graphics. *Free Batch Photo Resizer*. Černčice, 2005–2010. <<http://www.rw-designer.com/picture-resize>>.
- Roland, Martin. »Sto lat katalogowania średniowiecznych rękopisów iluminowanych w Wiedniu [A Century of Cataloguing Illuminated Manuscripts in Vienna]«. *Z badań nad książką i księgozbiorami historycznymi [Studies in the history of books and book-collections]* 3 (2009): 5–19. Englische Übersetzung online: <<http://paecht-archiv.univie.ac.at/dateien/cat-illum-mss-vienna.pdf>>.
- Roland, Martin. *Zur Wirkung von Katalogen illuminierter Handschriften*. 14th International Congress "Cultural Heritage and New Technologies", 2009 74–82. <http://www.stadtarchaeologie.at/wp-content/uploads/eBook_WS14_Part2_Workshops.pdf>.
- Roland, Martin. *Kurzinventar der illuminierten Handschriften bis 1600 in der Bibliothek des Zisterzienserstifts Stams in Tirol*. [Mit Beiträgen von Maria Theisen (zu Cod. 12) und Lilian Armstrong (zu Cod. 44). Konzeptuelle Gestaltung und technische Umsetzung Armand Tif. Wien: Bibliotheksstiftung Otto Pächt, 2010.] <<http://paecht-archiv.univie.ac.at/ki/stams.html>>.
- Schmidt, Gerhard. »Geleitwort«. Roland, Martin. *Buchschmuck in Lilienfelder Handschriften*. Wien: Niederösterreichisches Institut für Landeskunde, 1996. 7–8.
- Theisen, Maria. *HISTORY BUECH REIMENWEISZ. Geschichte, Bildprogramm und Illuminatoren des Willehalm-Codex Wenzels IV. von Böhmen*, Wien, Österreichische Nationalbibliothek Ser. Nov. 2643. Veröffentlichungen der Kommission für Schrift- und Buchwesen des Mittelalters, Reihe IV, Bd. 6, Wien: Österreichische Akademie der Wissenschaften, 2010.
- Tif, Armand. *Digitale kunsthistorische Inventarisierung mittelalterlicher Buchbestände im Web*. 14th International Congress "Cultural Heritage and New Technologies", 2009 67–73. <http://www.stadtarchaeologie.at/wp-content/uploads/eBook_WS14_Part2_Workshops.pdf>.
- Tif, Armand und Roland, Martin unter Mitarbeit von Maria Theisen und Alois Haidinger. *Kurzinventar der illuminierten Handschriften bis 1600 und der Inkunabeln in der Bibliothek des Augustiner-Chorherrenstiftes Herzogenburg in Niederösterreich*. [Wien: Bibliotheksstiftung Otto Pächt, 2009.] <<http://paecht-archiv.univie.ac.at/ki/herzogenburg.html>>.
- Tif, Armand. *Low-Budget-Konzept zur Online-Inventarisierung von Kleinsammlungen*. [Wien: Bibliotheksstiftung Otto Pächt, 2010.] <<http://paecht-archiv.univie.ac.at/ki/low-budget-konzept-zur-online-inventarisierung-von-kleinsammlungen.pdf>>.

Towards a Comparative Approach to Manuscript Study on the Web: the Case of the *Lancelot-Grail* Romance

Alison Stones, Ken Sochats

Abstract

This paper presents an outline of the on-going *Lancelot-Grail* Project, an interdisciplinary collaborative research project drawing together, analysing, and making available in text and picture the surviving manuscripts of the popular Arthurian romance known as the *Lancelot-Grail*. The project uses web technology as part of the analytical process and as a means to navigate within the material, presenting models based on the concepts of geographic information systems (GIS) in a non-traditional context.

Zusammenfassung

Der Beitrag stellt das laufende *Lancelot-Grail*-Projekt vor, das als interdisziplinäres und kollaboratives Forschungsprojekt Texte und Bilder der Handschriftenüberlieferung des unter dem Namen *Lancelot-Grail* bekannten populären Romans zusammenstellt, analysiert und online verfügbar macht. Das Projekt verwendet Web-Technologien sowohl für seine eigenen Analysen als auch als Navigationswerkzeug innerhalb des Materials. Es überträgt Modelle, die als Grundlage für Geoinformationssysteme (GIS) entwickelt wurden, in einen neuen Kontext.

1. Introduction

The *Lancelot-Grail* is the most popular version of Arthurian romance, surviving complete or in part in some 200 manuscript copies made between c. 1220 and 1504 (Woledge 1954 71-79, 1975 50-59). These manuscripts are housed today in libraries all over Europe and the USA but are available in the original only to scholars. Whereas some can be consulted complete in digitized form on the Bibliothèque nationale de France (BnF) site Gallica or on Mandragore on Digital Scriptorium (illustrated pages only), on the French provincial libraries site Enluminures, or on the sites of individual libraries,¹ many more are available only as selected illustrations in secondary literature.

¹ On Gallica: BnF fr. 95, 344, 16999; Mandragore: BnF fr. 105, 111, 113-116, 117-120, 122, 1422-1424, 9123, 19162, 24394; on the French provincial libraries site Enluminures: Tours, BM 961; Le Mans MM 354; Dijon

Some manuscripts are fully illustrated (London, BL Add. 10292–4 has 748 illustrations, the most of all), while others contain a single picture at the beginning of the major textual subdivisions or Branches into which the lengthy episodic narrative is divided (*L’Estoire del saint Graal*, *Merlin*, *Suite vulgate du Merlin*, *Lancelot* (with its own subdivisions), *La Queste del saint Graal*, *La Mort Artu*). Though attributed to authors Robert de Boron (*Estoire*, *Merlin*) and Gautier Map (*Queste*, *Mort Artu*), most of the manuscripts were made by anonymous scribes, decorators and illuminators. Some were made for or acquired by famous collectors—Jean de Berry († 1416), Jacques d’Armagnac († 1477)—but most patrons are unknown. We aim to determine what kinds of people found these texts interesting, where, and what aspects of the text the patrons and makers found compelling.

Why the *Lancelot-Grail*? In a previous research project organized by Alison Stones, a team of 15 specialists in literature, palaeography, codicology and history of art drew together an illustrated catalogue and essays on the 45 manuscripts and fragments of the romances of Chrétien de Troyes (see bibliography). In the project described in this paper we aim to make the manuscripts of the *Lancelot-Grail* and our findings available not just in print but also on the web, and to use web-based technology in the analytical and presentation processes.

The *Lancelot-Grail* Project began as a multi-institutional cross-disciplinary collaboration in the late 1990s, based at the University of Pittsburgh.²

2. Goals and Methods

Our aim is to enable on-line navigation of the manuscripts of the *Lancelot-Grail* romance and their illustrations in several ways, both synchronic and diachronic. We explore how the spatial analysis based on GIS concepts can be used in non-traditional applications, treating the manuscript page as a conceptual map in which different levels of information may be overlaid, using Active Server Pages to move within an individual folio, from a folio to an episode, and from an episode to a branch—or vice-versa. In the absence of a medieval term we define an ‘episode’ as a sub-section of a ‘branch’, a sequence of text and picture which concentrates on a particular hero or event (such as the False

BM 527: Digital Scriptorium: University of California, Berkley, Bancroft Library 106, 107; individual library sites: The John Rylands University Library of Manchester, French 1; Cologny-Genève, The Bodmer Library MS 147; New Haven, Yale University, Beinecke 229.

² Technical collaborators: Ken Sochats (Information Science and Telecommunications, University of Pittsburgh); Guoray Cai (Information Science, Pennsylvania State University), research assistant Jane Vadnal (Pittsburgh); Medieval French: †Elspeth Kennedy (Oxford); Medieval French, History of the Book and Ownership: Roger Middleton (Nottingham); Medieval French and Codicology: Keith Busby (Wisconsin); Medieval Art History: Alison Stones (Pittsburgh), Martine Meuwese (Leiden); graduate students Katherine Dimitrova, Marion Dolan, Julia Finch, Courtney Long, Kathryn Martin, Karen Webb (Pittsburgh) and Irène Fabry (Paris-III); Palaeographical Consultant: Michael Gullick.

Guinevere in the *Lancelot*; Maritime Adventures in *Estoire* and *Queste*). The term 'branch' (branke, branche) is medieval: it is used in several manuscripts to distinguish the major subdivisions of the text (*Estoire*, *Merlin*, *Lancelot*, *Queste*, *Mort Artu*), and sometimes to mark subdivisions within the *Merlin* and the *Lancelot*. Branch divisions have been followed by modern text editors. What can be learned is how different copies of the same text differ from each other in wording, picture, page layout and the like, prompting investigation of what those differences mean.

3. Pilot Project

We selected three manuscripts from the same 'workshop' made in Northern France or Flanders c. 1310–1325: London, British Library Additional 10292–4 and Royal 14.E.III, and a third copy now divided among Amsterdam, Bibliotheca Philosophica Hermetica MS 1, Oxford, Bodleian Library MS Douce 215 and Manchester, The John Rylands University Library MS French 1. We chose these because BL Add. 10292–4 is the most fully illustrated of all; it is complete; it contains the date of 12 February 1316 (1317 new style) 'carved' on a tomb on f. 55 in BL Add. 10292; it was the basis for Sommer's 1907–1913 edition of the text.

This phase was funded by the National Endowment for the Humanities and by Visiting Fellowships for Alison Stones at All Souls and Magdalen Colleges Oxford and Corpus Christi College Cambridge. We worked from printouts from old black and white microfilms, from first-hand study of the originals, and from new colour slides shot by Alison Stones at Amsterdam or purchased from the libraries (British Library, Bodleian Library, Rylands Library). Pittsburgh students scanned the slides with a view to eventually devising a web site. At the time, digital images were not available but the scanned images gave us useful working copies, particularly for text pages. In subsequent phases we were able to acquire better images, as described below.

4. Product Models

To analyse the manuscript page as a conceptual map on which can be plotted picture (subject, treatment), decoration (minor and major initials, pen-flourished, champie or foliate initials), text (episodes and events, names of characters), layout (columns, lines) and margins (decoration, annotations, blemishes) we developed a taxonomy of descriptors based on the contents of single pages (layout, script, decoration, illustration, text, marginalia, notes and annotations, physical signs of use and wear and tear, together with subdivisions of these categories). This was presented at the New Directions in Medieval Manuscript Studies conference at Harvard organized by Derek Pearsall in 1998 (Stones 2000). Each of the mark-up instances has associated spatial and descriptive

data further describing that particular mark-up. By using a standard descriptive framework, we were able to show how components of one page might be linked to similar components on other pages or in other manuscripts.

We built authority lists for the subjects of the illustrations of *Estoire*, *Agravain* (the last part of the *Lancelot*), *Queste* and *Mort Artu*.³

5. Developing a Series of Navigational Options

Navigating comparatively by ‘branch’ and ‘episode’ (cluster of related scenes) gives new insight in the use of the manuscripts.

Table 1 compares British Library (BL) manuscript Add. 10292 with Amsterdam, Bibliotheca Philosophica Hermetica (BPH) ms. 1, ii. BL Add. 10292 will be posted on the web later through DIAMM (see below). Eventually the images will be included in the comparative table and each folio will be linked to the folio analysis.

In this excerpt, BL Add. 10292 gives many more illustrations than Amsterdam although the text of the latter is not abbreviated. However the pictures in Amsterdam often include more detail than those in BL Add. 10292 where the action is played out over more scenes. A major difference is that BL Add. 10292 emphasizes the legal aspects of the challenge to Queen Guinevere’s legitimacy and the downfall of the lying False Guinevere and her sponsor. BL Add. incorporates a significant detail borrowed from legal illustration (cf. Gratian’s *Decretum*) where the motif of a lighted taper is used to indicate the pronouncement of excommunication and anathema. The patron of BL Add. is likely to have had a particular interest in the legal aspects of the text, unlike the patron of Amsterdam.

6. Extending the Parameters of the Project

This phase was funded by the Fulbright Foundation and by a Digital Innovation Fellowship from the American Council of Learned Societies for Alison Stones. Our goal was to obtain better quality images of the Pilot Project manuscripts and to broaden the scope of the project to include more manuscripts and to plot them across time and space.

High resolution scans from the British Library of the illustrated pages in BL Add. 10292–4 and Royal 14.E.III. were purchased.

Photography is done by the Digital Image Archive of Medieval Music (DIAMM).⁴ We had the illustrated pages in Amsterdam BPH 1 and Manchester, The John Rylands

³ See the appendices in Stones 1977, 1988, 2000, 2008, 2009, forthcoming; Blackman 1999, Meuwese 1999, Shailor 1984.

⁴ Cf. the chapter by Julia Craig-McFeely in this volume 307–339.

S IV 69. 17	<p>BL Add. 10292, f. 151r. Rubric: <i>Ensi que li rois Artus tint la roine Genieure par le main et la bailla a garder le roi Galehot par deuant lor baronie. et les barons en orent grant pitie.</i> King Arthur entrusts Queen Guinevere to King Galehot, witnessed by the barons. L, champie initial Text: <i>Lors prent li roys la royne et en uait a Galehot et li liure par le main...</i></p>	
S IV 72. 9	<p>BL Add. 10292, f. 152r. Rubric: <i>Ensi que on gete sentence sour le roy Artu.</i> The pope enjoins King Arthur to leave his new wife and take back the old one, extending towards him a lighted taper, a motif derived from the canon law indictment of excommunication and the pronouncement of anathema. O, champie initial Text: <i>Oor [sic] dist li contes que ensi est li roys Arthus departis de sa femme par le desloiaute de lautre Genieure...</i></p>	<p>Amst., BPH 1, ii, f. 227v. Rubric: <i>Chi gist malade de meselerie li fausse royne et li roys Artus le vint veir.</i> The False Guinevere, crowned, lying on her deathbed with Bertolai at her side holding a covered vessel, repeats her confession to King Arthur and the barons. O, pen-flourished initial Text: <i>Or dist li contes que ensi est li rois Artus partis de sa femme par le desloiaute de lautre Genieure...</i></p>
S IV 75. 16	<p>BL Add. 10292, f. 153r. Rubric: <i>Ensi comme li rois Artu se fist confesser dun hermite en son hermitage.</i> King Arthur, repentant, confessing to the hermit Amustans in his church-like hermitage. T, champie initial Text: <i>Tant dist mesires Gauvain au roy Artus son oncle...</i></p>	
S IV 76. 37	<p>BL Add. 10292, f. 153v. Rubric: <i>Ensi que li roys Artu et se baronie oient messe en.j. hermitage</i> King Arthur and his men hear mass in an elaborate, Gothic, hermitage. Q, champie initial Text: <i>Quant li roys ot ensi parler lermite si giete vn souspir...</i></p>	
S IV 79. 4	<p>BL Add. 10292, f. 154r. Rubric: <i>Ensi que li fause Genieure gist mesele et si uo (uo expunged) vint li rois Artu parler a li.</i> The False Guinevere, with veiled head, lying on her deathbed with Bertolai at her side holding a ciborium with a cross on top, repeats her confession to King Arthur and the barons. D, champie initial Text: <i>Dame vous gisies en si dolereuse carire comme cele qui a tout le pooir du cors perdu...</i></p>	

Table 1. The end of the False Guinevere episode, comparing BL Add. 10292 and Amsterdam BPH 1, ii (BL Royal 14.E.III lacks the Lancelot section of the text); text references are to Sommer (vol. 4); contractions have been silently expanded.

Lancelot-Graal Project
Amsterdam, Bibliotheca Philosophica Hermetica, MS 1, ii, f. 202r

Comparative Table	Screen-sized Image	Previous Folio	Previous Illustrated Folio	Next Folio	Next Illustrated Folio
-------------------	--------------------	----------------	----------------------------	------------	------------------------

[Rubric](#)

[Chapter Number](#)

Miniature

Champie initial

Text passage

Tear and repair

Erasure

Blemish

Text Passage

Chapter number

Penflourished initial and border

Blemish

erasure

© Courtesy of Amsterdam, Bibliotheca Philosophica Hermetica

Figure 1. Amsterdam, BPH, ms. 1, ii, f. 202r (opening of the False Guinevere episode from the *Lancelot* branch reproduced courtesy of the Bibliotheca Philosophica Hermetica, Amsterdam).

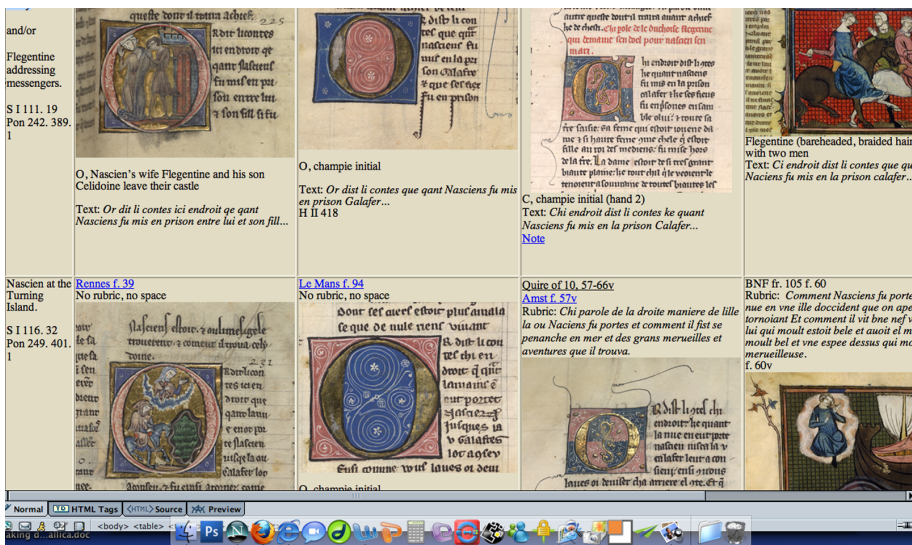


Figure 2. Maritime Adventures.

University Library French 1, Rennes BM 255 (c. 1220) and Le Mans MM 354 (c. 1285) photographed at high resolution by DIAMM.

We obtained permission from the Bibliothèque nationale de France to download for free and incorporate, or to link, *Lancelot-Grail* manuscripts from BnF sites Gallica and Mandragore. Downloading and resizing were done by University of Pittsburgh Art History graduate students.

We obtained permission to post images from the Bibliotheca Philosophica Hermetica, Amsterdam, and from Rennes BM 155 and Le Mans MM 354. The John Rylands University Library of Manchester has posted the images shot for the *Lancelot-Grail* Project by DIAMM on its web site. Lengthy attempts to negotiate a contract between the University of Pittsburgh and the British Library resulted in failure. We are now planning to display BL images through DIAMM under the terms of DIAMM's copyright agreement with the BL.

7. Comparative Selection and Treatment of Branches, Episodes, Folios

Fig. 2 shows a section from a comparative page of the Maritime Adventures episode from *Estoire* manuscripts in Rennes, Le Mans, Amsterdam and Paris (BnF). The choice

of champie initials links Le Mans and Amsterdam whereas long rubrics link Amsterdam and BnF and narrative scenes link Rennes and BnF: these findings point to levels of complexity in the transmission and illustration of these four copies which can be corroborated with further comparisons in order to reconstruct overall patterns of transmission and reception.

After the second phase of the *Lancelot-Grail* Project the collected data allowed the following conclusions: The manuscripts vary substantially in selection, placing, and treatment of illustrations. Sometimes a champie initial is substituted for an illustration. Of special interest is the treatment of Solomon's enchanted ship and its cross: sometimes as a cross on the sail, at other times a cross held by Tout-en-Tout in the ship. These pictorial variants depend on textual variants in the respective manuscripts, indicating that illustrators (or planners) paid careful attention to textual description (Stones 2009). A selector model is under development which will allow other manuscripts, branches, and episodes to be selectively compared.

8. The *Lancelot-Grail* Manuscript Tradition Across Time and Space

Using our expert art-historical and palaeographical analyses we plotted the chronological and geographical distribution of as many *Lancelot-Grail* manuscripts as possible, based on intensive research at the Bibliothèque nationale de France and on scattered manuscripts in many other collections (such as New York, Yale, Berkeley, Bonn and others).

On the project web site we linked manuscripts either to pages created by ourselves or to existing sites of individual libraries. The laborious task of downloading, resizing, labelling, has been done by graduate students in art history at the University of Pittsburgh. The task continues.

The interest in the *Lancelot-Grail* romance spread from Northern France in the early 13th century to England and Italy: the map in figure 3 shows this graphically and is a sample of a series of projected maps which will chart in 50-year increments the gradual spread of interest as reflected in patterns of collecting and gift-giving between c. 1220 and 1504, the date of the latest dated manuscript.

The goals of this phase of the project are first to use mapping to better identify where and when the manuscripts were manufactured, and then to relate differences in manuscripts to local political, cultural, economic and other conditions. We hope our plotting of changing patterns of interest in these stories may lead to correlations with the significant improvements made in measurement, navigation and other geographic technologies.

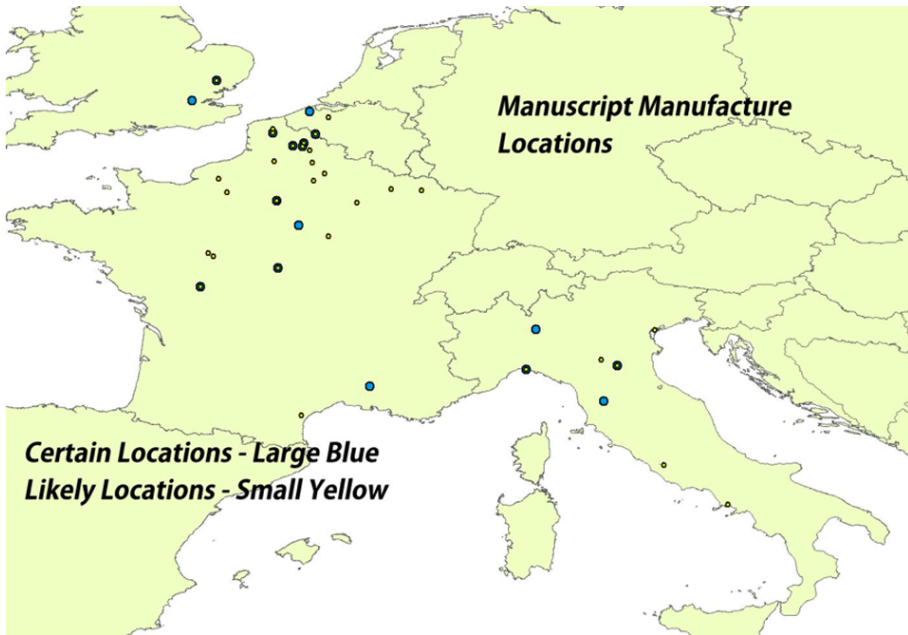


Figure 3. Sample map of distribution of *Lancelot-Grail* manuscripts.

9. Future Plans

The above outline indicates some of the directions that research on the *Lancelot-Grail* manuscripts have taken in the previous phases of research and development. The results of each phase can now be applied and exploited in more detail both on the web and in conference papers and published articles. Our goals are to make much more of our analysis available on the web and to unite in tabular form the illustrations with the descriptive and analytical research that has been carried out so far in the project. We hope our approach will be transferable to other manuscript projects.

Bibliography⁵

Beazley, Charles Raymond. *The Dawn of Modern Geography*. 3 vols. London: John Murray. 1897–1906.

⁵ Further references to articles by project participants are listed on the project web site.

- Blackman, Susan. "A Pictorial Synopsis of Arthurian Episodes for Jacques d'Armagnac, Duke of Nemours". *Word and Image in Arthurian Literature*. Ed. Keith Busby. New York-London: Garland, 1996. 3–57.
- Digital Scriptorium. Columbia: Columbia University Libraries, [1996–2010].
<<http://www.scriptorium.columbia.edu/>>.
- Edson, Evelyn. *Mapping Time and Space: How Medieval Mapmakers Viewed Their World. The British Library Studies in Map History*. Volume I. London: The British Library, 1997.
- The History of Cartography*. Eds. John Brian Harley and David Woodward. Chicago: University of Chicago Press, 1987.
- Kennedy, Elspeth and Alison Stones. "Signs and Symbols in the *Estoire del saint Graal* and the *Queste del saint Graal*." *Signs and Symbols* (Harlaxton Medieval Studies XVIII). Eds. John Cherry and Ann Payne. Donington (Lincs.): Shaun Tyas. 2009. 150–167.
- Lancelot-Graail Project*. Pittsburgh: University of Pittsburgh, [1995–2010].
<<http://www.lancelot-project.pitt.edu/>>.
- Les Manuscrits de Chrétien de Troyes*. Eds. Keith Busby et al. 2 Volumes. Athens (GA) / Amsterdam: Rodopi, 1993.
- Meuwese, Martine. "Three Illustrated *Prose Lancelots* from the same Atelier." *Text and Image: Studies in the French Illustrated Book from the Middle Ages to the Present Day* (Bulletin of the John Rylands University Library of Manchester 81/3). Eds. David J. Adams and Adrian Armstrong. Manchester: The John Rylands University Library, 1999. 97–125.
- Sommer, Heinrich Oskar. *The Vulgate Version of Arthurian Romances*. 7 vols. Washington (DC): Carnegie Institution, 1909–1913.
- Stones, Alison. "The Earliest Illustrated *Prose Lancelot* Manuscript?" *Reading Medieval Studies* 3 (1977): 3–44.
- Stones, Alison. "Some Aspects of Arthur's Death in Medieval Art." *The Passing of Arthur*. Eds. Christopher Baswell and William Sharpe. New York: Garland, 1988. 52–101.
- Stones, Alison. "Teaching and Research on the Web: Three Sites." *Computing and Visual culture: Representation and Interpretation*. [Fourteenth Annual CHArt Conference.] Ed. Tanya Szrajber. London: CHArt, 1999. 111–122.
- Stones, Alison. "The *Lancelot-Graail* Project." *New Directions in Later Medieval Manuscripts*. Ed. Derek Pearsall. York Medieval Studies. Woodbridge: Brewer. 2000. 167–82.
- Stones, Alison. "The Illustrations of *Mort Artu* in Yale 229: Formats, Choices, and Comparisons." *The Mort Artu in Yale 229*. Ed. Elizabeth Willingham. Turnhout: Brepols, 2008. 263–316.
- Stones, Alison. "The Illustrations of the *Queste del saint Graal* in Yale 229 and other *Queste* manuscripts." *The Queste del saint Graal in Yale 229*. Ed. Elizabeth Willingham. Turnhout: Brepols, forthcoming.
- Wolledge, Brian. *Bibliographie des romans et nouvelles en prose française*. Geneva: Droz, 1954. And *Supplément*. Geneva: Droz, 1975.

Artefacts and Errors: Acknowledging Issues of Representation in the Digital Imaging of Ancient Texts

Melissa M. Terras

The shadows of artefacts would constitute the only reality people in this situation would recognize.

Plato, the Republic, Book 7, 515c.

Abstract

It is assumed, in palaeography, papyrology and epigraphy, that a certain amount of uncertainty is inherent in the reading of damaged and abraded texts. Yet we have not really grappled with the fact that, nowadays, as many scholars tend to deal with digital images of texts, rather than handling the texts themselves, the procedures for creating digital images of texts can insert further uncertainty into the representation of the text created. Technical distortions can lead to the unintentional introduction of ‘artefacts’ into images, which can have an effect on the resulting representation. If we cannot trust our digital surrogates of texts, can we trust the readings from them? How do scholars acknowledge the quality of digitised images of texts? Furthermore, this leads us to the type of discussions of representation that have been present in Classical texts since Plato: digitisation can be considered as an alternative form of representation, bringing to the modern debate of the use of digital technology in Classics the familiar theories of mimesis (imitation) and ekphrasis (description): the conversion of visual evidence into explicit descriptions of that information, stored in computer files in distinct linguistic terms, with all the difficulties of conversion understood in the ekphratic process. The community has not yet considered what becoming dependent on digital texts means for the field, both in practical and theoretical terms. Issues of quality, copying, representation, and substance should be part of our dialogue when we consult digital surrogates of documentary material, yet we are just constructing understandings of what it means to rely on virtual representations of artefacts. It is necessary to relate our understandings of uncertainty in palaeography and epigraphy to our understanding of the mechanics of visualization employed by digital imaging techniques, if we are to fully understand the impact that these will have.

Zusammenfassung

Die paläographische, papyrologische und epigraphische Forschung geht davon aus, dass das Lesen eines beschädigten und radierten Textes Unsicherheiten mit sich bringt. Bislang haben wir uns jedoch noch nicht ausreichend mit dem Umstand auseinandergesetzt, dass heutzutage, wo sich viele Forscher eher mit digitalen Bildern von Texten als mit dem Text selber befassen, die Erstellung von digitalen Bildern ebenso zu Unsicherheiten in der Repräsentation des Textes führen kann. Technisch bedingte Verzerrungen können zu unbeabsichtigten 'Artefakten' in den Bildern führen, die sich auch in der Textrepräsentation niederschlagen. Wenn wir schon den digitalen Surrogaten der Texte nicht trauen können, können wir es dann ihren Transkriptionen? Wie gehen Wissenschaftler mit der Qualität digitaler Bilder von Text um? Das führt zu der Diskussion über die Repräsentation von Texten, die in der Altphilologie seit Plato geführt wird: Digitalisierung kann als eine alternative Form der Repräsentation gelten, welche die klassische Diskussion um Mimesis (Imitation) und Ekphrasis (Beschreibung) in die moderne Diskussion über die Nutzung digitaler Technologien einbringt. Die Umwandlung von visueller Evidenz in explizite Beschreibung von Information, gespeichert in einem Computer in voneinander klar getrennten sprachlichen Ausdrücken, zeigt all die Schwierigkeit von Umwandlungen, die als Ekphrasis verstanden worden sind. Die Fachgemeinschaft hat bislang noch nicht darüber nachgedacht, was es für ihr Forschungsfeld, praktisch wie theoretisch, bedeutet, von digitalen Bildern abhängig zu sein. Fragen nach Qualität, Kopie, Repräsentanz und Substanz sollten Teil unseres Dialoges werden, wenn wir digitale Surrogate dokumentarischen Materials benutzen, auch wenn wir ein Verständnis von dem, was es heißt, sich auf virtuelle Repräsentationen von Artefakten zu verlassen, gerade erst konstruieren. Wir müssen also unsere Vorstellung von Unsicherheit in Paläographie und Epigraphik mit einer Vorstellung von der Mechanik der Visualisierung mit digitalen Bildgebungsverfahren ergänzen, wenn wir denn ihre gesamten Auswirkungen auf die Forschung verstehen wollen.

1. Introduction

Constructing readings of ancient documents is a difficult, complex, and time-consuming task, often involving reference to a variety of linguistic and archaeological data sets, and the invocation of previous knowledge of similar documentary material. Due to the involved reading process, it is difficult to record how the final interpretation of the document was reached, and which competing hypotheses were presented, adopted, or discarded in the process. It is also difficult to acknowledge and present the probabilities, and uncertainties, which were called on to resolve a final reading of a text. It is assumed,

across all aspects of palaeography, codicology, papyrology, and epigraphy¹ (which share a central core of identifying aspects and making sense of documentary material, despite their individual focus on media or form) that a certain amount of uncertainty is inherent in the reading of damaged and abraded texts. Indeed, the Leiden system (and its related markup technique, EpiDoc) allows for the encapsulation of this uncertainty, and acknowledge that uncertainty is a critical part of the reading of ancient texts. Yet we have not really grappled with the fact that, nowadays, as many scholars tend to deal with digital images of texts, rather than handling the texts themselves, the procedures for creating digital images of texts, or relying on computational systems to aid the process of reading ancient texts, can insert further uncertainty into the representation of the text created. This is becoming a general problem in all fields of manuscript studies, and requires further focus and concentration.

In technical terms, issues raised by the digitisation process include distortion caused by lens shape, difficulties in colour management and reproduction, and the unintentional introduction of ‘artefacts’ into images, which can have an effect on the resulting image. If we cannot trust our means of reproduction of images of texts, can we trust the readings from them? How do scholars acknowledge the quality of digitised images of texts?

Furthermore, this leads us to the type of discussions of representation that have been present in Classical texts since Plato: digitisation can be considered as an alternative form of representation, bringing to the modern debate of the use of digital technology in Classics the well trodden arguments of mimesis (imitation and representation of aspects of the real world in different forms and media). Digitisation is, even, a form of ekphrasis, “a descriptive speech which brings the thing shown vividly before the eyes”² converting visual evidence into explicit descriptions of that information, stored in computer files in distinct linguistic terms, with all the difficulties of conversion understood in the process of ekphrasis.

Has the classical community considered what becoming dependent on digital texts means for the field? Are these issues of quality, copying, representation, and substance part of our dialogue when we consult digital images of ancient texts? What measures can be taken to ensure that we understand what we are looking at when utilising digital images of ancient texts? How does the form of the digital image inform or distort

¹ Whilst papyrology is the study of ancient manuscripts, mostly written on papyrus, palaeography is the study of handwriting, and the decipherment and reading of historical manuscripts. Epigraphy deals with Ancient inscriptions, whilst codicology is the study of books, especially manuscripts, as physical objects. All have a fundamental core function of extrapolating meaning and understanding about culture and history from written historical primary sources. Although the aims and foci of each speciality remain distinct, there are various crossovers in methodology and approach which allow for generalisations regarding the study of ancient texts and our approach to uncertainty therein.

² This definition appears in a series of first century CE Progymnasmatia from Theon, via Hermogenes. See Webb.

our potential readings of content, and relate to our understandings of uncertainty in palaeography and epigraphy? This chapter aims to explore issues of representation and uncertainty within the reading of ancient texts, with a particular focus on our increasing use of digital images, and how this impacts the papyrology and epigraphical community.

2. Papyrology and Computing

Classics as a subject has made much use of information technology (see Crane for an overview). “This tendency can partly be explained with reference to two observations: (1) the complexity of the textual, historical, linguistic, material, and artistic sources that need to be considered in classical scholarship, and (2) the patchy coverage and fragmentary state of many of these same artefacts” (Bodard and Mahony). Most of the uses of computing when reading ancient texts, however, do not turn to advanced computational techniques: like many disciplines in the humanities, the use of computing in Classics is mostly to speed up and enhance access to information which had previously only available in analogue format, through a process of digitising existing resources and making them available online. Deegan and Tanner summarise succinctly the wide range of reasons given for digitisation, and the advantages digitisation of materials can provide for scholars and institutions:

immediate access to high-demand and frequently used items; easier access to individual components within items (e.g. articles within journals); rapid access to materials held remotely; the ability to reinstate out of print materials; the potential to display materials that are in inaccessible formats, for instance, large volumes, or maps; ‘virtual reunification’—allowing dispersed collections to be brought together; the ability to enhance digital images in terms of size, sharpness, colour contrast, noise reduction, etc.; the potential to conserve fragile/precious objects while presenting surrogates in more accessible forms; the potential for integration into teaching materials; enhanced searchability, including full text; integration of digital media (images, sounds, video, etc.); the ability to satisfy requests for surrogates (photocopies, photographic prints, slides, etc.); reducing the burden of cost of delivery; the potential for presenting a critical mass of materials. (32–33)

Classical scholars were swift to recognise the benefits of digitisation, and began to address major issues of computational infrastructure in the 1970s with large, fairly centralized efforts which have since become the central starting point for many aspects of the scholarly research of antiquity (Crane). These projects included David Packard’s Ibycus system, the *Thesaurus Linguae Graecae*, the *Database of Classical Bibliography*,

the *Bryn Mawr Classical Review*, the Duke Databank of Documentary Papyri, and the Perseus Project. By the close of the 20th Century, many online projects devoted to delivering images, transcriptions, and notes regarding papyri had been set up. A project such as APIS, the Advanced Papyrological Information System (2007a), demonstrates how digitisation, and the drawing together of disparate existing knowledge sources, can aid those engaged in reading ancient Texts. APIS describes itself as

a collections-based repository hosting information about and images of papyrological materials (e.g. papyri, ostraca, wood tablets, etc) located in collections around the world. It contains physical descriptions and bibliographic information about the papyri and other written materials, as well as digital images and English translations of many of these texts. When possible, links are also provided to the original language texts (e.g. through the Duke Data Bank of Documentary Papyri). The user can move back and forth among text, translation, bibliography, description, and image (APIS 2007c).

APIS currently hosts over 30,000 different records, and 23,000 individual images from 27 major institutions (providing online links to digital images, where available, at their host site if they are not part of the APIS system themselves). With the provision of an intuitive search function, scholars can search across a wealth of texts, and can often have access to various high resolution images of each document, allowing them to download them and access them remotely, negating the need to travel, visit, and handle the document themselves.

Likewise, sites such as the Vindolanda Tablets Online (n.d.) devoted to the documents found at one particular fort on Hadrian's Wall, provide much greater access to both images, transcriptions, translations, and notes regarding documents, allowing cross referencing and in depth scholarly analysis of one particular set of documents to be undertaken without having to gain access to the original artefacts in the British Museum nor depend on print volumes in which photographic provision can sometimes be limited. A recently developed companion site, Vindolanda Tablets Online II provides updates to the collection, and provides a web service where the existing information concerning the tablets can be searched in finer detail than in the original site. The Vindolanda websites are based on print volumes, with the online equivalent extending their functionality and increasing the volume of information available regarding texts. Other sites (such as the Inscriptions of Aphrodisias project, or Inscriptions of Roman Tripolitania) have adopted the online medium to extend and expand the remit of the print editions they are based on, using a sophisticated blend of content specific markup (EpiDoc) behind the interface to allow searching, and scholarly analysis, both in and across the individual collections.

It is not the place here to survey every website which provides online versions of print volumes of ancient texts, or high resolution images of papyri, ostraca, epigraphical

and other documents.³ The point to be made is that in a short space of time, much work in papyrology has moved from consideration of the physical document itself, or print surrogates (which, although previously commonly used have their own limitations in print and image quality) to working with relatively high resolution images of the documents, provided by online scholarly editions through databases. When Roger Bagnall, in his seminal 1995 introductory text “Reading Papyri, Writing Ancient History” opened with the phrase “Papyrology has tended to one of the most resolutely technical and positivistic disciplines of antiquity” (vii) he was not concerned with web based technologies which were then in their infancy. Likewise, when H. C. Youtie commented in a lecture in 1957 that

the distinguishing characteristics of the scholarship of the twentieth century has been its dependence on papyri, and papyrology, like archaeology, epigraphy, numismatics, and mediaeval palaeography, has become a permanent adjunct to the technical equipment of classical scholars and ancient historians (267)

he was not concerning himself with Information Technology. As Brunner said in a 1993 chapter covering “Classics and the Computer, the history of a relationship”:

The field of Classics encompasses quite a few centuries; yet there are few distinct periods within this history that can be said to have witnessed changes as rapid and fundamental as through brought about by the entry of the field into the electronic world. A mere two and a half decades after “electronic machines” first found mention in an AP publication, few (if any) members of the Association remain uninvolved in, and unaffected by, computing. (28)

Papyrology is a field which revolves around the resolution and consideration of uncertainty within ancient texts. We may acknowledge this in textual transcriptions of documentary material, but do we acknowledge how the dependence on digital surrogates can affect and inject other modes of ambiguity, uncertainty, and representation in the images of documents that we now attempt to read?

3. Uncertainty, Ancient Texts, and Computing

The process of reading ancient documents is traditionally undertaken by an expert such as an epigrapher, papyrologist, or palaeographer. The expert will use their accumulated knowledge combined with external resources to piece together an interpretation of each

³ The Digital Classicist Wiki hosts a list of projects undertaking relevant research, including those who provide digitised surrogates of papyrological material. A list of papyri sources is also given in Bernhardt.

ancient document. Such interpretation can be a long process, and it can be difficult for experts to maintain a record of the decisions made whilst undertaking their reading (Youtie 1963). This is important when defending their interpretation, sharing their hypotheses with other experts, or breaking off in the process of reading an ancient text and hoping to pick up the thought process in another reading session.

When undertaking tasks which depend on complex reasoning, there is often an element of probability that needs to be addressed. Unlike mathematical models which can deal with, say, different levels of dosage of medication on a patient, those reading ancient documents are often faced with a range of uncertainties with few or no prior models on which to base reasoning, and little framework in which to test these hypotheses. Ambiguity is a feature of the readings of ancient texts, even in their published version. This is not a critique of papyrologists: published versions are open to correction, and merely detail the extent to which the author has resolved the reading of the text at that moment in time. Bowman and Tomlin provide examples of readings which have changed dramatically between different versions of published texts. Tomlin demonstrated how the correct reading of a tablet was achieved by rotating the tablet through 180 degrees and rereading the text. However, the fact that this is accepted practice indicates that uncertainty about the reading of ancient text is seldom exhausted. Reading resolves around the resolution of ambiguity through prediction, prior knowledge, reasoning about the characteristics of the documents, and the head-on addressing of uncertainties.

For over a century, those reading ancient texts have tried to encapsulate their reasoning process in the resulting published transcripts of the texts in question. Due to the costs involved, producing a facsimile of a text proved prohibitive (Mahoney), and scholars became dependent on the use of set of signs and brackets in transcriptions to signify textual features such as lost or supplied characters, damage to the text, the expansion of abbreviations, etc. Grenfell and Hunt were consistent in their use of signs and brackets to describe the state and reading of texts from Oxyrhynchus, influencing a generation of scholars. Bidez and Drachmann published a pamphlet examining the different customs of using bracket and sign conventions in editions. Van Groningen used this analysis to suggest a unified system for marking editions (1932a). However, the International Congress of Orientalists agreed upon a different system which strongly resembles the papyrological system used by Grenfell and Hunt. It is this system that is now known as the Leiden System (Van Groningen 1932b) which aims to capture various characteristics of a text, widely adopted in print publications for all types of ancient texts.

A commonly used symbol within the Leiden system is the under-dot, used to represent uncertainty. However, this is one of the most confusing concepts to represent in such a transcription. There is no way to measure the extent of uncertainty (for example, where the reader is a little trepidatious that their interpretation is correct, or marking

that the letter in question is unreadable). Practice varies between papyrologists, with some using the underdot to represent a broken letter which is certain, others using it only when identification is in doubt. The authors of the *Tabulae Vindolandesens* I and II (Bowman and Thomas 1983, 1994) tend to use it both when identification is uncertain and when to show that there is no doubt: using the dot to show that the letter is broken. An analysis of the Leiden Markup of the Vindolanda ink text indicates that 9.9% of the letters in the ink text were marked as being broken (Terras 71): the identification of a proportion of these will be problematic. Uncertainty is therefore an important issue to address when building computational systems to aid papyrologists. Unfortunately, encapsulating uncertainty in computational systems is not a straightforward process.

Computational systems depend on resolving “real world” situations into exact numerical strings. The ordinary, or “real” world of our senses, exists in a continuous flowing stream of signals across time and often space. An ancient document—or a photograph of the document—exists in analogue, where a varying signal represents a continuous range of values. In order to record, copy, transmit, or analyse such a complex signal using computational power, it is necessary to translate this into a form which is more simple, predictable, and processable. All telecommunication systems have one thing in common: the information to be sent is converted into signals which can be transmitted, and reassembled on reception, to be converted into something we can perceive as a fair copy of the original. Digital systems are those which rely on a sequence of discreet numeric values, rather than the unconstrained and continually varying qualities of analogue signals. Numeric values are used in digital systems for processing, display, transmission, and input: often sampling values from analogue sources in a process called “digitisation”.

The most common digital systems are those used in computing and electronics, which rely on the binary numeric system. This is a system which represents all numbers using only two symbols, typically 0 and 1. These zeroes and ones are known as binary digits, or more commonly as the shortened derivation: “bits”. Strings of bits can represent text, images, sound, and moving images: as the information to be represented grows more complex, more bits are required to represent it, and more complex mechanisms are used to store, display, and process the information contained within the data stream.

Providing numeric, textual, image, sound, and video based data in digital format, whether they have been translated from an analogue signal into bits, or “born digital” by being created with computational technologies in the first place, has various advantages. These strings of bits can be easily replicated, transmitted, accessed, and processed. Saving the data in a structured, predetermined format means it may be device independent, and can be transferred from system to system with minimal problems. Data can be manipulated by dedicated computer programs, allowing new versions of the information to be generated. Data can also be processed: mathematically sorted through to show hidden relationships, new arrangements, different views, and

expanded, contracted or concatenated knowledge. Human eyes and ears can sometimes distinguish between continuous analogue signals, and bit by bit digital approximations. Digital media are at their most effective when their constituent parts—samples—are not detectable by human senses.

The digitized images of ancient texts delivered by internet technologies to our desktops therefore have a more complex relationship to their original manifestation than we may like to consider, given that the visualization reproduced on the screen is often so seductive:

A digital representation of an artefact is a representation of certain relevant characteristics of the artefact. It is not the original and complete artefact, nor even a metonymy or simulacrum of the complete artefact. It is only a representation of some “relevant characteristics”. (Arnold 127)

Further to just producing realistic representations of ancient texts, the development of new imaging techniques such as multi spectral imaging, image processing algorithms which can remove noise, character recognition tools which can propose combinations of strokes as possible letters, and tools which can search databases to match word fragments to relevant words, grammar, and orthographies, depend on building explicit representations of knowledge which computational systems can work with. Mathematics underpins all image processing and Artificial Intelligence systems, depending on concrete algorithms and defined representations of information. The humanities scholar’s grasp of uncertainty when approaching damaged and abraded texts, written in a foreign, ancient language, is not something that maps easily onto computational systems (or our expectations of the results computational systems can produce). It is hard to provide enough real-world knowledge encapsulated computationally which can reflect the amount of contextual information necessary to undertake the reading of ancient texts (Terras). Additionally, many of these relatively new and emergent technologies which supply encouraging results and novel readings of documents, such as multi spectral imaging as applied to ancient texts (see Ware et al., Bearman), are not benchmarked and thoroughly tested to ensure their methodologies are robust, and that the new images created from the process have a mathematically sound relationship to the original artefact.

4. Quality and Quality Assurance in Digitised Collections

If scholars are using digital surrogates to produce readings of ancient texts, it is imperative that these images are of high quality. But what makes a “good” digital image of an ancient document? How do we know the digital representations that we

work with are fit for purpose? How can we assess image quality, and trust the resulting representation?

The creator of the image has the purpose of communicating it through a suitable channel to one or more observers. Every element of the chain from creator to receiver affects the quality of the image, and hence the effectiveness of the communication process. Noise in the channel, however, may degrade the transmission, and the characteristics of both the origination medium and reproduction medium limit the overall ability of the system to render the image (size, density range, colour gamut, etc). Viewing conditions in each case affect the perception of the person viewing the image. (Macdonald and Jacobsen 352)

Many digitization projects either produce their own, or adhere to established, guidelines for the production of digital images. APIS, for example, produce their own guidelines for contributing organizations which details file format, resolution, the use of colour targets to allow calibration of colour, lighting, and even file naming (APIS 2006). The Library of Congress produces up to date “Technical Standards for Digital Conversion of Text and Graphic Materials” (Library of Congress), which many digitization projects consult (and should be the first port of call for those considering a digitisation project, as they provide the minimum, and best practice, specification for digitisation processes). In the UK, the Joint Information System’s Committee, which provides leadership in IT services for Higher Education, provides a free advisory service, JISC Digital Media, for those creating and dealing with digital material (2010). Given that “few standards govern the creation and use of digital images, and in a world of multiple stakeholders and multiple perspectives it may be difficult to agree on a uniform approach that suits all circumstances” (Kenney 24) it is imperative that those undertaking digitisation programs consult guidelines and carry out benchmarking procedures to ensure quality control of the digitised output (see Kenney for an overview).

However, it can be difficult to assess the quality of digital images themselves, even if there are assurances from the creators that guidelines have been followed. Many things are lost in the sampling of an image to create a digital representation: it can be difficult to ascertain size, physical characteristics, texture, and the accuracy of colour. Although certain measures can be taken to ensure the capture process is as accurate as possible, including adequate documentation (see MacDonald and Jacobsen), many projects do not comply with these. For example, the images within the Vindolanda Tablets Online do not have a colour target supplied with them, meaning that scholars could not calibrate their monitors to best look at the images—if they indeed understood why and how they should do so. Some, but not all, of the images presented in *Inscriptions of Aphrodisias* are supplied with a measurement bar—partly because many of the images were decades old before digitisation—and this can lead to confusion about scale.

It is good practice to inform users of the digitisation process, and how and when the primary material was digitised. Vindolanda Tablets Online does this (2003), as does Inscriptions of Aphrodisias (Bodard and Spence), although the documentation requires some searching for on both of their websites. It can be difficult within APIS to see where and how the digital record, and digital image, of each individual papyrus was created, as this technical metadata is not recorded about each entry in the database (APIS 2007b). This lack of easy access to documentation regarding the digitisation process can create problems for scholars who are dependent on the outputs of digitisation projects to undertake their own original research, as studies have shown that digital humanities resources tend to be poorly documented and therefore not be trusted by users:

In the absence of technical documentation, it was impossible to reuse files [...] Although users require procedural documentation, about the status and completeness of sources, and selection methods, this is often difficult to locate [...] and shows that this makes reuse of digital resources almost impossible. (Warwick et al. 33)

Furthermore, the persuasive nature of the visualisation and display can mean we do not stop to question the very nature of digital images of historical artefacts:

its labor of production has been concealed and therefore bears less evidence of authorship, provenance, originality, and other commonly accepted characteristics attributed to physical objects. For these reasons the digital object's materiality is not well understood. (Cameron 70)

There is a lot that can go wrong in the creation of digital images that are used for representation of historical artefacts. Colour, in particular, is a thorny issue, with many issues such as illumination, and the differences between the way the human eye and computer systems record colour (see Hunt for an introduction). There are technical issues with the sampling and representation process that depend on the nature of device characteristics that determine spatial resolution, and effects that are created dependent on the type of camera lens used for image capture (see Holst for an overview). Quantifying the accuracy of imaging equipment—whether analogue or digital—is not a simple mathematical task (Keelan). Assessing image quality is usually dependent on human observation, and this is a subjective notion dependent on observer-based quality judgements (Engeldrum). Additionally, conversion from one format to another, sampling and resampling, and compressing file sizes can result in unintended visual effects, or “artefacts” becoming obvious in images, where information is deleted and inserted awkwardly into the resulting representation.

It is rare that those utilizing digital images for scholarly research would stop to consider the mechanics which produced that image, or their accuracy or veracity. It

is difficult, even with adequate documentation, for those undertaking digitization to express the capture conditions and relationship of the surrogate to the digitized item in question. Images may be from a trusted scholarly source—but did those capturing the images understand and comply with the technical issues necessary to best represent that ancient object in the modern digital realm?

5. Digitisation and Theory

It may be useful to pause and consider the theoretical frameworks that exist for dealing with surrogates, or representations, of the objects that we wish to study. The reading of ancient texts when faced with image based source material is an ekphrastic task: that is, in general, papyrology attempts to produce a faithful textual description of image based material. Ekphrasis, “the verbal representation of visual representation” (Heffernan 297) can be traced back to the legendary Shield of Achilles in the *Iliad*, and stems from ancient poetics and rhetoric. The impossibility of describing the visual within the textual realm is at once both recognised, and discounted by the “ekphrastic hope” (Mitchell 152) where the impossibility of ekphrasis is overcome in imagination, metaphor, and sense discovery from the image itself. The difference between text and image is both celebrated and feared, acknowledged and discounted: can the description of an image in textual form ever match up to the visual sensory perception of seeing that image in the flesh?

The utopian figures of the image and its textual rendering as transparent windows onto reality are supplanted by the notion of an image as a deceitful illusion, a magical technique that threatens to fixate the poet and the listener. (Mitchell 156)

Ekphrasis can be used as a tool to focus “the interarticulation of perceptual, semiotic, and social contradictions within verbal representation” (180). The literature regarding ekphrasis is large, with each type of visual representation “such as photography, maps, diagrams, movies, theatrical spectacles [... carrying] its own peculiar sort of textuality into the heart of the visual image” (181). Can images of ancient texts, and the description of the texts we visualise therein, be discounted from the problems understood and articulated by the task of describing, accurately, in text the visual nature of image based material?

Furthermore, digital images can be viewed as fundamentally ekphrastic: digital image data itself generally consists of a list of values of the colours of individual pixels, and associated instructions for computing applications regarding how this data should be displayed, written in a data stream in a computer file. This draws us back again to the notion of representation in the digital realm, and the reduction of our sensory

experience back to zeroes and ones. What is lost in this ekphrastic translation? What accuracy is compromised through this point by point sampling, and expression into explicit numerical (if not textual) values? Our images of text become numbers: we translate them back to images, then to transcriptions. Are we careful enough that nothing is changed or lost in the process?

Digital images are also surrogates, and have a complex relationship to their analogue equivalent which theorists are only now beginning to start to question:

The digital historical object can exist in many realms and perform many roles that go beyond representation, interpretation, education, documentation, and archive. Indeed its analogonic role is potentially diverse [...] the status of copies from nondigital originals still remains ambiguous [...] A range of expanded meaning, material characteristics, and behaviours emerge as representing a particular configuration of space, time, and surface, sequence of user activities—a particular formal material and user experience. (Cameron 68)⁴

It is tempting to embrace the value of material authenticity and adopt a repugnant stance to reproductions of primary historical material. Both Walter Benjamin and Jean Baudrillard have argued that mechanical reproduction and simulations pose a threat to the real object of focus, leading to the loss of its “auraic, iconic, and ritualistic qualities” (Cameron 50). Do students currently learning palaeographic methods focus on digital surrogates? Is there a need to ensure that the “Google generation” (JISC 2007) who will be the professors of papyrology of the future understand the relationship the surrogate has to the original? Have they even handled original texts? Does this at all matter? In his chapter in this volume Peter Stokes remarks specifically on this subject

basic skills in handling original materials, in reading, transcribing, editing and understanding these objects are central to [textual] studies. The question that remains is therefore twofold. First, how can digital tools be used to better teach traditional skills. Second, a question much less frequently raised, is how the teaching of traditional skills should or could itself change as a result: how and to what extent should digital content be explicitly introduced into the curriculum for the study of [...] manuscripts? (Stokes)

It is of course nonsensical to suggest that the benefits brought by digitisation to the palaeographical community should be rejected for some notion of the material superiority of original documentary materials. However, it may be helpful to give ourselves some distance from the digital image of ancient texts, and consider the

⁴ This whole paper provides a good overview of available theoretical discussions regarding the relationship of the digital surrogates to their analogue counterpart.

implications of our dependency on digital images for scholarly research. By doing so, we may reach an understanding of the “modality and materiality of digital historical objects” as “new roles and a set of defining characteristics emerge beyond their role as servant to the ‘real’ as representation, presence, affect, experience, and value” (Cameron 70).

6. Dependency on the Digital: When Surrogates become Primary Sources

It is worth saying here that although it is a possibility, I am not aware of any published documentary material that has been read erroneously (and published, and refuted) due to faults in the digitization process. I am not aware of any digitised versions of ancient texts that are so faulty that artefacts and errors within the digital image are obvious⁵. I am aware that the *intention* with most cultural and heritage digitisation projects is to provide the *best* digital representation of that object that is possible. The intention is to replace the need for travel across countries and continents to see a scrap of papyrus the size of your hand, and to facilitate research. I do not mean to criticise the efforts of those undertaking digitization (nor the particular projects named above).

It is also worth saying that the problems in ensuring that we understand the nature of digital representations of primary sources also apply to other disciplines. Any historical or literary research which depends on image-based primary resources is now facing the same turning point, as scholars turn to the digital as a convenient means to view and access a wide variety of digitised content. Some issues will hit fields harder than others (colour reproduction in digital images, for example, should generally be more of an issue for the art historian than the papyrologist), but nevertheless, the problem is similar: scholars become trained in the tools and methodologies entrenched in their own discipline, but seldom are educated in the technical underpinnings which allow the primary sources they depend on from various points across the world to magically appear on the computer screen in their office. Understanding digitisation is then a particular extension of digital literacy:

the confident and critical use of Information Society Technology (IST) for work, leisure and communication. It is underpinned by basic skills in ICT: the use of computers to retrieve, assess, store, produce, present and exchange information, and to communicate and participate in collaborative networks via the Internet (European Parliament and Council)

combined with information literacy:

⁵ Or, in the case of Google Book’s digital image of the first page of a Victorian edition of Plato’s *Euthyphron*, where the fingers of the digitization operative are clearly visible in the scan (Cohen).

knowing when and why you need information, where to find it, and how to evaluate, use and communicate it in an ethical manner. (CILIP)

Those utilising digital image resources of primary textual material have an obligation to their academic discipline to understand the nature of the resources they are basing readings, transcriptions, and translations on. Additionally, scholars should be more open when publishing readings of texts in articulating both the sources and methodologies used when consulting digital resources, and their use of digital surrogates in scholarship. This will have the added benefit of providing “evidence of value” of costly digital resources, which are battling at the moment to prove that they are essential services to academic communities, and deserve further funding (AHRC). Likewise, those providing digital images of ancient Texts, or any historical documents, have an obligation to fully document, describe, and elucidate the process by which the digital surrogates were created. Scholars must be encouraged to use this documentation, to ensure that their research is based on as authentic a representation as a primary text as possible: otherwise the readings they generate from them simply cannot be trusted.

7. Conclusion

It would be folly to suggest that we should return to pre-digital dependencies on the physical document and print based surrogates when trying to transcribe, read, study, and understand ancient documentary material. The affordances of digital media increase productivity, reduce travel time and cost, and provide in-depth and detailed information regarding individual texts. Vast collections of images of ancient texts are accessible to scholars from their own desk—and although few make this explicit in the methodologies published in research papers and monographs, many are now dependent on online databases and databanks.

However, just as palaeographers, epigraphers, papyrologists and codicologists are educated and trained in textual mores of the ancient world, those scholars dependent on the digital environment should ensure that they understand the representations of artefacts that they base their research upon. Those undertaking digitisation projects should be aware of the minimum acceptable technical standards and adequate documentary approaches for the digital representations created (see above for references). Those producing digital surrogates of primary historical texts should produce adequate documentation that is easily available regarding the technical procedures involved in capturing images of the artefacts. Only by fore-fronting the use of both standards and documentation of these standards can we produce robust digital resources that can stand up to academic scrutiny.

However, there are issues regarding digitisation that cannot be resolved in practical form, and we must begin to build up our theoretical understanding of notions of

digitisation and representation so we can articulate our dependencies and be sure about our methodologies when relying on digital surrogates. In particular, it should be acknowledged that digital images of ancient texts have a complex relationship to their source material. Additionally, digital images created in ways which would never have existed using traditional photography, or human vision, such as multi spectral or infra-red images, should be treated as they are: representations, and surrogates, rather than replicas of original documentary material. Furthermore, more thought should be given as to the computational representational structures that we shoehorn our understanding of image, text, and language into, when experiencing the convenience of online papyrological sources.

The aim of this chapter has been to raise issues of digital representation within the Classics, and particularly within the papyrology community. The rapid computational transformation that has occurred in the field must be followed by the questioning and inquisitive methodology which is applied to trying to understand ancient texts themselves: what does it mean to continually use digital image surrogates to produce readings of ancient texts. If we cannot understand the means of production of the surrogates, can our interpretations ever be robust? Only through becoming digitally and informationally literate can we trust that our images of artefacts are free from artefacts and errors.

Bibliography

- APIS. *APIS Imaging Standards*. 2006. <http://www.columbia.edu/cu/lweb/projects/digital/apis/imaging_guidelines_2006Jan.pdf>.
- Advanced Papyrological Information System. 2007a. <<http://www.columbia.edu/cu/lweb/projects/digital/apis/>>.
- Guidelines for APIS Metadata Contributors. 2007b. <<http://www.columbia.edu/cu/libraries/inside/projects/apis/guidelines.html>>.
- About the APIS project. 2007c. <<http://www.columbia.edu/cu/libraries/inside/projects/apis/about.html>>.
- Arnold, David. "Digital Artefacts. Possibilities and Purpose." *The Virtual Representation of the Past*. Eds. Mark Greengrass and Lorna Hughes. Ashgate: Farnham. 2008 158–170.
- ICT in Arts and Humanities Research. Review Activities*. Arts and Humanities Research Council. Reading: University of Reading, 2007. <<http://www.ahrcict.rdg.ac.uk/activities/review/index.htm>>.
- Baudrillard, Jean. *Simulacra and Simulation*. Trans. Sheila Faria Glaser. Ann Arbor: University of Michigan Press, 1994.
- Bagnall, Roger S. *Reading Papyri, Writing Ancient History*. London and New York: Routledge, 1995.
- Bearman, Gregory. "Imaging the Dead Sea Scrolls for conservation purposes." *SPIE Newsroom*. 29th December 2008. <<http://spie.org/x32760.xml?ArticleID=x32760>>

- Benjamin, Walter. "The Work of Art in the Age of Mechanical Reproduction." *Illuminations: Essays and reflections*. Trans. Harry Zohn. London: Jonathan Cape, 1970 (1st published 1935). 217–252.
- Bernhardt, Theodore. *Papyri Pages, Site Links*. 2009. <<http://papyri.tripod.com/links-2.html#Collections>>.
- Bidez, Joseph and Anders B. Drachmann. *Emploi des signes critiques, disposition de l'apparat dans les éditions savantes de textes grecs et latins, conseils et recommandations*. Paris: É. Champion for Union Académique International, 1932.
- Bodard, Gabriel and Simon Mahony. "Though much is taken, much abides': Recovering antiquity through innovative digital methodologies: Introduction to the special issue." *Digital Medievalist* 4 (2008). <<http://www.digitalmedievalist.org/journal/4/DCintro/>>.
- Bodard, Gabriel and Paul Spence. "Technical Preface." *Aphrodisias in Late Antiquity*. London: King's College London, 2004. <<http://insaph.kcl.ac.uk/ala2004/about/techpref.html>>.
- Bowman, Alan and J. David Thomas. *Vindolanda: The Latin Writing Tablets*. London: Society for Promotion of Roman Studies, 1983.
- Bowman, Alan and J. David Thomas. *The Vindolanda Writing-Tablets (Tabulae Vindolandenses II)*. London: British Museum Press, 1994.
- Bowman, Alan K. and Rodger S. O. Tomlin. "Wooden Stylus Tablets from Roman Britain." *Images and Artefacts of the Ancient World*. Eds. Alan K. Bowman and Michael Brady. Oxford: Oxford University Press, 2005. 7–14.
- Brunner, Theodore. F. "Classics and the Computer: The History of a Relationship. *Accessing Antiquity: The Computerization of Classical Databases*. Ed. Jon Solomon. Tucson: University of Arizona Press. 1993. 10–33
- Cameron, Fiona. "Beyond the Cult of the Replicant: Museums and Historical Digital Objects – Traditional Concerns, New Discourses." *Theorizing Digital Cultural Heritage, A Critical Discourse*. Eds. Fiona Cameron and Sarah Kenderdine. Cambridge MA: MIT Press, 2007.
- CILIP. "Information Literacy: Definition." London: Chartered Institute of Library and Information Professionals (CILIP), 2010. <<http://www.cilip.org.uk/get-involved/advocacy/learning/information-literacy/pages/definition.aspx>>.
- Cohen, Dan. "Google Fingers." *Blog Post, Dan Cohen's Blog*, 26th June 2006. <http://www.dancohen.org/blog/posts/google_fingers>.
- Crane, Gregory. "Classics and the Computer: An End of the History." *A Companion to Digital Humanities*. Eds. Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell, 2006. <<http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-4&toc.depth=1&toc.id=ss1-2-4&brand=default>>.
- Creating Digital Images*. Excerpt from "Introduction", Alan K. Bowman and J. David Thomas. *The Vindolanda Writing Tablets*. Volume III. London: British Museum Press, 2003: 14. Online: <<http://vindolanda.csad.ox.ac.uk/tablets/TVdigital-2.shtml>>.
- Deegan, Marilyn and Simon Tanner. *Digital Futures: Strategies for the Information Age*. London: Library Association Publishing, 2002.
- Digital Classicist Projects. London: Centre for Computing in the Humanities, King's College London – Kentucky: Stoa Consortium, University of Kentucky, 2008. <<http://wiki.digitalclassicist.org/Category:Projects>>.

- Engeldrum, Peter.G. *Psychometric Scaling*. Winchester MA: Imcotek Press, 2000.
- European Parliament and Council. Recommendation on key competences for lifelong learning. Brussels: European Union, 2005.
<http://ec.europa.eu/education/policies/2010/doc/keyrec_en.pdf>
- Grenfell, Bernard. P. and Arthur. S. Hunt. *The Oxyrhynchus Papyri: Part I*. London: Egypt Exploration, 1898.
- Heffernan, James. "Ekphrasis and Representation." *New Literary History*, 22.2 (1991): 297–316.
- Holst, Gerald C. *CCD Arrays, Cameras and Displays*. 2nd Edition. Winter Park et al.: SPIE Press, 1998.
- Hunt, Robert G. W. *The Reproduction of Colour*. Hoboken, NJ: Wiley, 2004.
- Inscriptions of Aphrodisias*. Eds. Charlotte Roueché et al. London: King's College, 2007.
<<http://insaph.kcl.ac.uk/iaph2007/index.html>>.
- Inscriptions of Roman Tripolitania*. Eds. J. M. Reynolds and J. B. Ward-Perkins. London: King's College 2009. <<http://irt.kcl.ac.uk/irt2009/index.html>>.
- JISC. *Google Generation*. Eds. David Nicholas et al., London: JISC, 2007.
<<http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/googlegen.aspx>>.
- JISC Digital Media*. [Bristol: University of Bristol] 2010. <<http://www.jiscdigitalmedia.ac.uk/>>.
- Keelan, Brian W. *Handbook of Image Quality*. New York: Marcel Dekker, 2002.
- Kenney, Anne R. "Digital Benchmarking for Conversion and Access." *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Eds. Anne R. Kenney and Oya Y. Rieger. Mountain View (CA): Research Libraries Group, 2000. 24–60.
- Library of Congress. *Technical Standards for Digital Conversion of Text and Graphic Materials*. Washington (DC): Library of Congress, 2007.
<<http://memory.loc.gov/ammem/about/techStandards.pdf>>.
- MacDonald, Lindsay and Ralph Jacobsen. "Assessing Image Quality." *Digital Heritage, Applying Digital Imaging to Cultural Heritage*. Ed. Lindsay MacDonald. Oxford: Elsevier, 2006. 351–373.
- Mahoney, Anne. "Epigraphy." *Electronic Textual Editing*. Eds. Lou Burnard, Katherine O'Brien O'Keeffe and John Unsworth. New York: Modern Language Association. 2006. 224–240.
- Mitchell, W. J. Thomas. *Picture Theory*. Chicago: University of Chicago Press, 1994.
- Stokes, Peter. "Teaching Manuscripts in the Digital Age" *Codicology and Palaeography in the Digital Age II*. Eds. Franz Fischer, Christiane Fritze and Georg Vogeler. Norderstedt: Books on Demand 2010. 229–245.
- Tawse, Anne. *The Digitisation of Medieval Manuscripts, Success or Failure?* Unpublished MA Dissertation. London: School of Library, Archives and Information Studies, University College London, 2003.
- Terras, Melissa. *Image to Interpretation: Intelligent Systems to Aid Historians in the Reading of the Vindolanda Texts*. Oxford: Oxford University Press, 2006.
- Tomlin, Roger S. O. "Vinisius to Nigra: Evidence from Oxford of Christianity in Roman Britain." *Zeitschrift fur Papyrologie und Epigraphik (ZPE)* 100 (1994): 93–108.
- Van Groningen, Bernhard Abraham. "Projet d'unification des systemes de signes critiques." *Chronique d'Egypte* 7 (1932a). 262–269.

- Van Groningen, Bernhard Abraham. "De signis criticis in edendo adhibendis." *Mnemosyne* 594 (1932b): 362–365.
- Vindolanda Tablets Online*. Ed. Centre for the Study of Ancient Documents et al. (n.d.). <<http://vindolanda.csad.ox.ac.uk/>>.
- Vindolanda Tablets Online II*. Ed. Centre for the Study of Ancient Documents and The British Museum [Oxford: University of Oxford], 2010. <<http://vto2.csad.ox.ac.uk/>>.
- Ware, Gene et al. "Multispectral Document Enhancement: Ancient Carbonized Scrolls." *Geoscience and Remote Sensing Symposium. 2000. Proceedings. IGARSS 2000. IEEE 2000 International* Vol. 6 (2000): 2486–2488.
- Webb, Ruth. "The Progymnasmata as Practice." *Education in Greek and Roman Antiquity*. Ed. Yun Lee Too. Leiden: Brill, 2001: 289–316.
- Warwick, Claire et al. "Documentation and the Users of Digital Resources in the Humanities." *Journal of Documentation* 65.1 (2009): 33–57.
- Youtie, Herbert C. "The Papyrologist: Artificer of Fact." *Greek, Roman and Byzantine Studies* 4 (1963): 19–32.
- Youtie, Herbert. C. "Papyrology and the Classics." *Zeitschrift für Papyrologie und Epigraphik* (ZPE) 67 (1957): 267–281.

Digitaler Katalog und Semantik



Digital Catalogue and Semantics

Handschriften, Nachlässe, Inkunabeln & Co. – Die Erschließung der deutschen Handschriften und die Bereitstellung von Sonderbeständen in Online-Katalogen an der Universitätsbibliothek Tübingen mit TUSTEP

Silke Schöttle, Ulrike Mehringer

Zusammenfassung

Die deutschen Handschriftenbestände der Universitätsbibliothek Tübingen wurden von 2007 bis 2010 neu erschlossen und für die Forschung in einem Online-Katalog zur Verfügung gestellt. Die Arbeitsumgebung für die Kurzerschließung der knapp 2200 Signaturen ist ein auf die Bedürfnisse des Bestandes zugeschnittenes Makro des Tübinger Systems von Textverarbeitungsprogrammen (TUSTEP), mit dem die Universitätsbibliothek bereits weitere Kataloge für bibliothekarische Sonderbestände, insbesondere für Nachlässe und Inkunabeln, realisieren konnte. TUSTEP hat sich während des Erschließungsworkflows insbesondere durch seine komfortablen Anzeige-, Filter- und Suchmodi sowie durch seine Flexibilität bei der Satzerstellung für einen Bandkatalog mit verschiedenen Registern und der gleichzeitigen Möglichkeit der Bereitstellung eines Online-Katalogs mit zahlreichen Recherchemöglichkeiten bewährt.

Abstract

The German manuscripts of the Tübingen University Library were newly catalogued between 2007 and 2010 and made accessible online. The work to prepare short descriptions for around 2,200 manuscripts was realised using a macro of the Tübingen system of word processing programs (TUSTEP), which had been adapted for the purposes of this project. The macro had already been used for the cataloguing of other special collections such as literary estates and incunabula. TUSTEP has been useful especially through its comfortable modes for searching, filtering, and display during cataloguing as well as its flexibility during the typesetting of a printed catalogue with different indices and its possibility for online publication of the data.

1. Die Tübinger Handschriftenbestände

Die Universitätsbibliothek Tübingen kann derzeit fast 2230 von circa 5800 Handschriften, rund 250 Nachlässe und über 2100 Inkunabeln als Sonderbestände in speziellen elektronischen Katalogen außerhalb des Tübinger Bibliothekskatalogs (OPAC) nachweisen. In den letzten drei Jahren wurde ein solcher Online-Katalog für die deutschen Handschriften der Universitätsbibliothek erarbeitet und damit ein vielfältiger und für die Tübinger Universitäts-, Wissenschafts- und Kulturgeschichte wertvoller Bestand online recherchierbar gemacht. Der heterogene Bestand von überwiegend neuzeitlichen Handschriften des 16. bis 20. Jahrhunderts war bisher lediglich und teilweise mühsam über einen handschriftlichen Katalog des 19. Jahrhunderts zugänglich, der bis in die jüngste Zeit sukzessive fortgeführt worden war. Auf dessen Grundlage konnte der deutsche Handschriftenbestand jetzt mit kurzen Beschreibungen erschlossen werden, wofür die Stiftung Kulturgut Baden-Württemberg die notwendigen Mittel bereitstellte. Ziel des Projekts war von Beginn an ein Online-Katalog mit umfangreichen und flexiblen Recherchemöglichkeiten auf der Homepage der Universitätsbibliothek Tübingen.

Die Geschichte der Bestandsbildung haben Kehr (Handbuch 20–109) und Röcklein (Handschriftenkatalog 1991 11–49) ermittelt. Nachdem der Unterbibliothekar Jeremias David Reuss (1750–1837) Ende des 18. Jahrhunderts erstmals die Handschriften von den Drucken getrennt aufgestellt und um 1780 einen ersten Tübinger Handschriftenkatalog angelegt hatte, erhielt der gesamte Bestand in den Jahren 1839 bis 1841 seine noch heute existierende Ordnung. Der Philologieprofessor und Bibliothekar an der Universitätsbibliothek Adelbert von Keller (1812–1883) ordnete die Handschriften damals, seinem eigenen wissenschaftlichen Interessensgebiet entsprechend, in ein philologisch orientiertes System ein, das auf Sprachgruppen und einzelnen Sprachen basierte und innerhalb der Gesamtsystematik der Bibliothek das Signaturenkürzel M mit weiter unterteilenden Kleinbuchstaben (Ma–Mh) und römischen Ziffern erhielt. Der nach Sprachgruppen geordnete Bestand wurde dabei um verschiedene formalsachliche Teilbestände wie Autographen, Musikalien und Urkunden ergänzt (Mi–Ml). Ende des 20. Jahrhunderts fügte die Tübinger Handschriftenabteilung noch weitere Teilbestände für Fragmente (Mm) und Nachlässe (Mn) hinzu. Die umfangreichsten Handschriftenteilbestände neben den insgesamt weniger als 120 Signaturen umfassenden griechischen (Mb), germanischen (Me), romanischen (Mf) und slawischen (Mg) Handschriften stellen heute die lateinischen (Mc, 380 Signaturen), die orientalischen (Ma, 1863 Signaturen) und schließlich mit fast 2850 Signaturen die deutschen Handschriften mit den Württembergica (Md, Mh, Mh I, Mh II, Mh III) dar.

Die deutschen Handschriften, deren Großteil aus dem 18. und 19. Jahrhundert stammt und auf dem Weg der Schenkung oder des Erwerbs ganzer Gelehrtenbibliotheken aus dem Umfeld von Stadt und Universität Tübingen in die Universitätsbibliothek gelangte, stellen eine wertvolle, aber bisher nur wenig beachtete Quelle zur Tübinger

Ma	Orientalische Handschriften	1863
Mb	Griechische Handschriften	46
Mc	Lateinische Handschriften	380
Md	Deutsche Handschriften (mit den bis 1983 erworbenen Nachlässen)	1129
Me	Germanische Handschriften (außer deutsche)	17
Mf	Romanische Handschriften	53
Mg	Slavische Handschriften	2
Mh	Württembergische Handschriften	1717
Mi	Autographen	869
Mk	Musikalien	97
Ml	Urkunden	23
Mm	Fragmente	
Mn	Nachlässe (seit 1983 erworben)	

Tabelle 1. Die Handschriftensignaturengruppen der Universitätsbibliothek Tübingen mit Zahl der Handschriften je Gruppe.

Universitätsgeschichte dar. Gleichzeitig findet sich darin viel Unvermutetes, das auch weit über die Grenzen der Universitätsstadt von Bedeutung und Interesse ist. Eine große Bandbreite an Disziplinen, Institutionen und Persönlichkeiten bestimmt die Inhalte und Überlieferungsformen der Handschriften. Juristen, Philologen, Historiker, Theologen, Mathematiker, Mediziner, Meteorologen, Numismatiker, Kartographen, Missionare und Reisende, Klöster und Kanzleien hinterließen nicht nur Monographien, sondern auch Urkunden und Akten, Amts- und Formularbücher, wissenschaftliche Texte, Materialsammlungen und Korrespondenz. Zum Bestand zählen Grammatiken, Chroniken, Tagebücher, Erbinventare, Predigten, Vorlesungsmitschriften, Skizzen- und Gebetbücher, Reisebeschreibungen, studentische Stammbücher und vieles mehr. Dieser Kleinkosmos, der so manche Kuriosität birgt, reicht beispielsweise von der württembergischen Chronik des 16. Jahrhunderts (Mh 6) über die Anleitung zum Bau einer Hebmaschine für die Entwurzelung von Bäumen aus dem 18. Jahrhundert (Md 134) bis zur Dokumentierung der Eröffnung des Tübinger anatomischen Instituts im Wintersemester 1835/1836 (Md 303 Bd. 57).

2. Die Handschriftenererschließung mit TUSTEP

Die Handschriften werden mit TUSTEP (Tübinger System von Textverarbeitungsprogrammen; s.a. ITUG; Schneider-Lastin) erschlossen, das sich in der Universitätsbiblio-

```

PROG*MD.TIT
24.1      &001
24.2      &002 25.05.2007
24.3      &003 04.07.2007
24.4      &005
24.5      &010 Md 24
24.6      &030 #/+Hippokomik%-e#/- <Abschrift>
24.7      &040
24.8      &050 92 Bl.
24.9      &060 34x22
24.10     &070 o.0.
24.11     &080 Ende 16. Jh.
24.12     &100*Grisone, Federico
24.13     &102c366
24.14     &104bBlifers, Julius
24.15     &106c72
24.16     &108vLangenmantel, Johann Christoph
24.17     &110c598
24.18     &112vAutenrieth, Christoph Friedrich
24.19     &114c35
24.20     &200
24.21     &400
24.22     &410 Halblederband der UBT (19. Jh., Blifers)
24.23     &420 Bearbeitung Johann Faysers der "Ordni di cavalcare" von Federico
           Grisone; Traktat über die Reitkunst, Zucht, Dressur und Abrichtung von
           Pferden
24.24     &430 Abschrift aus: Grisone, Federico: Hippokomike. Künstlicher Bericht Und
           lundl allerzierlichste beschreibung ... Friderici Grisonis ...: Wie die
           Streitbarn Pferdt ... zum Ernst und lundl Ritterlicher Kurtzweil. geschickt
           und vollkommen zumachen: Jn sechs Bücher bester Ordnung, voluerständlichem
           Teutsch, vnd zierlichen Figuren ... dermassen in druck verfertigt, das
           dergleichen in Teutschland niemals ersehen worden/ Durch Johann Faysen den
           Jüngern von Arnstain .... Augspurg 1573 (UBT: Ah U1 9.2)
24.25     &435 Zahlreiche mit Aquarellfarben kolorierte Federzeichnungen
24.26     &440 Buchbesitzer: Johann Christoph Langenmantel; Geschenk der Erben des
           Stallmeisters Christoph Friedrich Autenrieth (1838, August 8)
24.27     &500
24.28     &502 #k+Reuss#k-, Catalogus 1877
24.29     &510 #k+Keller#k-, Verzeichnis bemerkt: #/+Vorher mit Md 23
           zusammengebunden#/-
24.30     &600 Ugl. Md 12, Md 23
24.32     &630 Gebhardt, Walther: Spezialbestände in deutschen Bibliotheken. Berlin
           1977. S. 476
24.33     &640
24.34     &700 Pferdezucht; Hippologie
24.35     &750 http://www.inka.uni-tuebingen.de/daten/mssbilder/md24.jpg
25.0
5 Zshf. gefunden      5 Sätze gefunden      17474 durchsucht
*=24.1 Gib Anweisung >

```

Abbildung 1. Datensatz mit Feldcodierungen (Md 24).

thek Tübingen bereits bei der Erstellung der gedruckten Kataloge für die lateinischen und griechischen Handschriften durch Röckelein, Brinkhus, Mentzel-Reuters und Wilhelm sowie der Bereitstellung von Online-Katalogen für die Inkunabeln (seit 2000) und Nachlässe (seit 2001) bewährt hat. Friedrich Seck und Ulrike Mehringer entwickelten das Makro *HAND*, das eine exakt auf die deutschen Handschriftenbestände und ihren heterogenen Charakter abgestimmte Beschreibung mit flexiblen Anpassungsmöglichkeiten während des Workflows sowie zahlreiche Recherchezugänge ermöglicht. In dieses Makro sind von Anfang an zwei Darstellungsoptionen eingebaut worden: zum Einen die Option, die erschlossenen Handschriften über eine Internetpräsentation zugänglich zu machen, zum Anderen die Möglichkeit, eine Druckvorlage mit Registern zu erzeugen, wie Seck sie beschreibt.

http://www.inka.uni-tuebingen.de/hand.php

Tübinger Handschriftenkatalog **ubTÜBINGEN**

[Inhalt und Benutzungshinweise](#) [Stücksuche](#) [Personensuche](#) [Körperschaften](#) Programmiert mit TUSTEP

[Zurück zur Ergebnisliste](#) | [Neue Suche](#)

Signatur: Md 24
 Grisone, Federico: *Hippokomikē* (Abschrift)
 - 92 Bl. - 34×22 cm - o.O. - Ende 16. Jh.

Einband: Halblederband der UBT (19. Jh., Biffers)

Inhalt: Bearbeitung Johann Faysers des "Ordini di cavalcare" von Federico Grisone; Traktat über die Reitkunst, Zucht, Dressur und Abrichtung von Pferden
Abschrift/Exzerpt/Übersetzung aus: Abschrift aus: Grisone, Federico: *Hippokomikē*. Künstlicher Bericht Vnd [und] allerzierlichste beschreybung ... Friderici Grisonis ... Wie die Streibarn Pferd ... zum Ernst vnd [und] Ritterlicher Kurtzweil, geschickt und vollkommen zumachen. In sechs Bücher bester Ordnung, voluerständlichem Teutsch, vnd zierlichen Figuren ... dermassen in druck verfertigt, das dergleichen in Teuschland niemals ersehen worden/ Durch Johann Faysen den Jüngern von Arnstein ... Augspurg 1573 (UBT: Ah VI 9.2)

Illustrationen: Zahlreiche mit Aquarellfarben kolorierte Federzeichnungen

Provenienz: Buchbesitzer: Johann Christoph Langenmantel; Geschenk der Erben des Stallmeisters Christoph Friedrich Autenrieth (1838, August 8)

Alte Tübinger Signatur II: REUSS, Catalogus 1877

Alte Katalogeinträge: KELLER, Verzeichnis bemerkt: Vorher mit Md 23 zusammengebunden

Zugehörige Hss.: Vgl. Md 12, Md 23

Literatur: Gebhardt, Walther: Spezialbestände in deutschen Bibliotheken. Berlin 1977, S. 476

Abbildungen: <http://www.inka.uni-tuebingen.de/daten/mssbilder/md24.jpg>

Beteiligte Personen / Körperschaften:

- [Grisone, Federico](#) (Verf.)
- [Biffers, Julius](#) (Buchbinder)
- [Langenmantel, Johann Christoph](#) (Vorbesitzer)
- [Autenrieth, Christoph Friedrich](#) (Vorbesitzer)

[Zum Seitenanfang](#)

Universitätsbibliothek Tübingen




Abbildung 2. Ergebnisanzeige des Datensatzes im Tübinger Handschriftenkatalog (Md 24).

Die Handschriften werden in drei miteinander verknüpften Dateien erschlossen, einer Titeldatei, einer Personen- und einer Körperschaftsdatei. Die Titeldatei enthält die eigentliche Erschließung der Handschriften in Datensätzen, die sich aus den verschiedenen mit Feldcodes benannten und markierten Feldern mit möglichst exakt zugewiesenen Inhalten zusammensetzen. Die wichtigsten Erschließungskriterien und Felder der Kurzbeschreibung der deutschen Handschriften sind Signatur, Band- und Stückzählung, Titel, Umfang, Maße, Entstehungsort und Entstehungsdatum der Handschrift, Autoren und andere Personen (etwa Schreiber, Buchbinder, Illustratoren, Vorbesitzer), Körperschaften, nähere inhaltliche Erläuterungen, Einbanddaten, Provenienz, alte Signaturen, Literaturangaben, Sekundärformen und Registerschlagwörter. Weitere Felder stehen bereit für die Erfassung von Illustrationen, alten Katalogeinträgen, zugehörigen Handschriften, Verweisen auf Vorgängerkataloge, Erhaltungszuständen

oder möglichen Schäden sowie für die URL einer digitalen Abbildung oder einer weiterführenden Webressource. Einbändige und mehrbändige Handschriften sowie Sammelhandschriften werden durch die Anbringung entsprechender Parameter unterschieden. Auch die verschiedenen Funktionen der im Datensatz genannten Personen können durch Parameter direkt nach dem Feldcode gekennzeichnet werden und so in den Online-Katalog übertragen werden (z. B. * = Autor, b = Buchbinder, s = Schreiber).

Die mit der Titeldatei mitgeführten Personen- und Körperschaftsdateien beinhalten zentrale Daten der in der Titeldatei immer wieder erscheinenden Personen und Körperschaften, die bei ihrem Vorkommen dort lediglich durch eine entsprechende Identifikationsnummer mit den passenden Lebensdaten oder Kurzbiographien verknüpft werden können. Die Personen- und Körperschaftsnamen werden dabei nach den Regeln der alphabetischen Katalogisierung angesetzt und, wenn möglich, direkt aus der Personennamendatei (PND) und Gemeinsamen Körperschaftsdatei (GKD) übernommen. Insbesondere die Personendatei enthält darüber hinaus weitere Felder für Quellenangaben, Berufs-, Lebens- und Wirkungsdaten sowie für die Verweisungsformen der angesetzten Namen.

Die komfortablen Anzeige-, Filter- und Suchmodi in TUSTEP bilden ein äußerst flexibles Arbeitsinstrument, mit dem in den Handschriftenbeständen Zerstreutes, etwa Provenienzen, Exlibris, Einbanddaten oder Illustrationen virtuell zusammengeführt und ausgedruckt werden können. Die Umsetzung der miteinander verknüpften TUSTEP-Dateien ermöglicht zudem sehr breit und flexibel angelegte Rechercheoptionen im Online-Katalog, zu denen nicht nur die Suche über verschiedene Felder in einer Suchmaske, sondern auch eine der Signaturenfolge entsprechende Kurztitelliste mit der Möglichkeit zum Schmökern zählt. In Zukunft soll die Funktionalität des Katalogs durch eine noch in Arbeit befindliche Listenansicht der vergebenen Personen-, Orts-, Provenienzen- und Sachregisterschlagwörter ergänzt werden.

3. Die Erstellung des Online-Katalogs mit TUSTEP

Die Umsetzung der Datenbasis in den Online-Katalog wird ebenfalls mit TUSTEP vorgenommen. Für die Suchmaske müssen zunächst die Suchfelder bestimmt werden. Grundsätzlich wäre jedes erfasste und mit einem Feldcode benannte Feld recherchierbar. Der Übersichtlichkeit halber wird jedoch für die Suchmaske eine sinnvolle Auswahl der im Format definierten Felder getroffen. Diese Auswahl orientiert sich an den zu erwartenden Recherchestrategien der Nutzer, die bei Bedarf angepasst werden können. Ein Suchaspekt kann sich auch auf mehrere Felder erstrecken, wie es beispielsweise von der Freitextsuche über alle Felder und der Personensuche über Ansetzungs- und Verweisungsformen bekannt ist. Es gibt drei verschiedene Suchmasken, mit denen jeweils die Titeldatei, die Personendatei und die Körperschaftsdatei durchsucht werden. Die Verknüpfung der drei Dateien findet über die Ergebnisanzeige statt. Eine

Registerfunktion, um in der Titelsuche auch auf die Verweisungsformen aus der lokalen Personendatei zugreifen zu können, ist wünschenswert und in Planung. Das gilt ebenso für Indices der Titel- und Schlagwortsuche. Für die Ergebnisanzeige der Titel gibt es derzeit drei Ausgabeformate. In der Maske *Stücksuche* kann zwischen einer Kurztitelliste und der Anzeige mit Kurzbeschreibung gewählt werden. Von der Kurztitelliste ist über die Treffernummer ein Wechsel auf die Kurzbeschreibung möglich, von dort führt ein Link auf die ausführliche Beschreibung im Vollformat.

Die Eingaben aus der Suchmaske werden an ein cgi-Skript übergeben, das vor dem Aufruf der TUSTEP-Sitzung die nötigen Systemvariablen definiert. Die Standard-Ausgabe erfolgt UTF-8-codiert. Derzeit gibt es drei Makros für die drei Varianten *Stücksuche*, *Personensuche* und *Körperschaftssuche*. Nach der Definition der Umgebungsvariablen endet das cgi-Skript mit dem Start der TUSTEP-Sitzung und dem Aufruf des Makros. Die Eingaben aus dem Formular werden an das jeweilige TUSTEP-Makro übergeben. Damit der Transport von Umlauten und Sonderzeichen keine Fehler produziert, werden die Daten umkodiert. Dem Wesen nach handelt es sich bei dem Tübinger Handschriftenkatalog um eine Flat-File-Datenbank mit flacher Datenstruktur, die durch die Verwendung von mehreren Dateien eine erweiterte Dimension erhält.

Um mit den Formulareingaben eine Suche durchzuführen, muss zunächst definiert werden, wo die Datendatei liegt und welche Struktur sie hat. In der Makrofunktion STRUCTURE werden den Feldnamen Variablennamen für die Ausgabe zugeordnet. Die Struktur der Datendatei – also das Datenformat, bzw. Kategorienschema – muss zwar ein paar wenigen Grundregeln folgen, ist aber ansonsten beliebig und jederzeit erweiterbar. Die Daten werden vor der Bereitstellung auf dem Webserver einer Syntaxprüfung unterzogen. Unbekannte Feldcodes oder eine falsche Reihenfolge der Feldcodes produzieren eine Fehlermeldung und führen zum Abbruch der Recherche. Ein TUSTEP-Makro findet formale Fehler und gibt ein Protokoll aus, anhand dessen die Daten korrigiert werden können.

Für jedes Suchfeld aus der Maske wird in den Suchbedingungen einzeln festgelegt, wie die Eingabe behandelt werden soll. Im Handschriftenkatalog wird beispielsweise Groß- und Kleinschreibung nicht unterschieden und Leerzeichen werden wie andere Sonderzeichen als Wortgrenze interpretiert. Mehrere Wörter innerhalb eines Suchfeldes werden mit logischem UND verknüpft, wobei die Reihenfolge keine Rolle spielt. Dies hat zur Folge, dass ein fester Ausdruck mit mehreren Wörtern durch Anführungszeichen und eine Trunkierung durch die Platzhalter * (beliebig viele Zeichen) und ? (genau ein oder kein Zeichen) gekennzeichnet werden muss. Eine Ausnahme bildet die Suche nach der Signatur, bei der das Leerzeichen nicht als Wortgrenze gilt und übergangen wird. Für die Recherche nach Jahreszahlen ist auch eine Bereichssuche in der Form < 1600, > 1900 oder 1850-1855 möglich. Die gefundenen Datensätze werden in den Variablen der Struktur abgelegt und können so über die cgi-Schnittstelle in HTML in beliebiger Reihenfolge als Ergebnisliste angezeigt werden.

Eine spezielle Import-/Exportfunktion gibt es im Tübinger Handschriftenkatalog derzeit nicht. Der Datenimport wäre aber über TUSTEP möglich, in dem das Kommando #UMWANDLE die Daten zu TUSTEP-Daten umkodiert und die Feldstruktur mit einem individuell angepassten Makro umgesetzt wird. Im Inkunabelkatalog INKA wurde dies schon vielfach mit unterschiedlichsten Quelldaten realisiert. Für den Export können ASCII-Dateien mit einem vorher individuell festgelegten Kategorienschema erzeugt werden. Bisher wurde diese Möglichkeit im Rahmen des Inkunabelkatalogs für Datenrücklieferungen an den Gesamtkatalog der Wiegendrucke in Berlin und den Inkunabel-Census an der Bayerischen Staatsbibliothek in München angewendet. Zudem können jederzeit Postscript-Dateien als Druckvorlage für einen Papierausdruck des Katalogs mit Registern erzeugt werden. Zusätzliche Exportformate könnten bei Bedarf entwickelt werden.

Die abschließende Frage, ob TUSTEP eine geeignete Arbeitsumgebung für die Erschließung von Handschriften darstellt, lässt sich aufgrund der sehr guten Erfahrungen nicht nur in dem hier beschriebenen Projekt positiv beantworten: In den letzten Jahren wurden an der Universitätsbibliothek Tübingen mehrere in ihren Grundfunktionalitäten ähnliche, aber jeweils auf ihren besonderen Inhalt abgestimmte Online-Kataloge für Sonderbestände mit TUSTEP realisiert.

- *Inkunabelkatalog INKA* (2000): aus Quelldateien unterschiedlicher teilnehmender Bibliotheken; mit Nebendateien für Provenienzen und Einbandwerkstätten; Verknüpfungen zu GW und ISTC
- *Tübinger Nachlasskatalog* (2001): mit Sucheinschränkung nach Materialart; Detail- und Übersichtssuche
- *Tübinger Handschriftenkatalog* (2010): mit Nebendateien für Personen und Körperschaften
- *Orientalische Handschriften des Evangelischen Stifts Tübingen* (2010)
- *Stammbücher der Herzogin Anna Amalia Bibliothek Weimar* (im Aufbau): mit einer zusätzlichen Suchebene für Einträge
- *Nachlasskatalog Hermann Zapf der Stadtbibliothek Nürnberg* (im Aufbau)

Insgesamt stellt TUSTEP also ein auf die konkreten Bedürfnisse der Erschließung spezieller bibliothekarischer Sonderbestände anwendbares und äußerst anpassungsfähiges Arbeitsmittel dar. Das Programm bewährt sich sowohl während der Erschließungsphase als auch bei der anschließenden Satzerstellung insbesondere immer wieder durch flexible und individuelle, jederzeit auch austauschbare Möglichkeiten, die einem Baukastensystem gleichen.

Bibliographie

Inkunabelkatalog INKA. Tübingen: Universitätsbibliothek Tübingen, 2000–2010.

<<http://www.inka.uni-tuebingen.de/>>.

ITUG: *International TUSTEP User Group*. <<http://www.itug.de/>>.

Handbuch der historischen Buchbestände in Deutschland. Bd. 9. Hg. Wolfgang Kehr. Baden-Württemberg und Saarland T–Z. Hildesheim: Olms-Weidmann, 1994.

Handschriftenkataloge der Universitätsbibliothek Tübingen. Hg. Joachim-Felix Leonhard.

Bd. 1: Die lateinischen Handschriften. Teil 1: Signaturen Mc 1 bis Mc 150, beschrieben von Hedwig Röcklein unter Mitwirkung von Gerd Brinkhus, Harald Weigel und Ulrike Hascher-Burger unter Benutzung der Vorarbeiten von Eugen Neuscheler. Wiesbaden: Harrassowitz, 1991.

Bd. 1: Die lateinischen Handschriften. Teil 2: Signaturen Mc 151 bis Mc 379 sowie die lateinischen Handschriften bis 1600 aus den Signaturengruppen Mh, Mk und aus dem Druckschriftenbestand, beschrieben von Gerd Brinkhus und Arno Mentzel-Reuters. Wiesbaden: Harrassowitz, 2001.

Bd. 2: Die griechischen Handschriften der Universitätsbibliothek Tübingen. Sonderband Martin Crusius, Handschriftenverzeichnis und Bibliographie, bearbeitet von Thomas Wilhelmi. Wiesbaden: Harrassowitz, 2002.

Nachlasskatalog Hermann Zapf der Stadtbibliothek Nürnberg [im Aufbau]. Nürnberg: Stadtbibliothek Nürnberg, 2008. <<http://tustep.stadtbibliothek.nuernberg.de/>>.

Orientalische Handschriften des Evangelischen Stifts Tübingen. Tübingen: Universitätsbibliothek, 2010. <<http://www.inka.uni-tuebingen.de/stift.php>>.

Schneider-Lastin, Wolfram. *TUSTEP-Tutorial*. Zürich, 2008.

<<http://elbanet.ethz.ch/wikifarm/schneider-lastin/index.php?n=Main.TUSTEP-Tutorial>>.

Seck, Friedrich. *Die Tübinger Handschriftenkatalogisierung. Datenformat und Datenverarbeitung* [unveröffentl. Manuskript der Handschriftenabteilung der Universitätsbibliothek Tübingen, z. Zt. letzte Version vom 14.04.2010].

Stammbücher der Herzogin Anna Amalia Bibliothek Weimar [im Aufbau]. Hg. Herzogin Anna Amalia Bibliothek Weimar u. Klassik Stiftung Weimar. Tübingen: Universitätsbibliothek Tübingen, 2010. <<http://www.inka.uni-tuebingen.de/stamm.php>>.

Tübinger Handschriftenkatalog. Tübingen: Universitätsbibliothek Tübingen, 2010.

<<http://www.inka.uni-tuebingen.de/hand.php>>.

Tübinger Nachlasskatalog. Tübingen: Universitätsbibliothek Tübingen, 2001.

<<http://www.inka.uni-tuebingen.de/nachlass.php>>.

Tübinger System von Textverarbeitungs-Programmen. TUSTEP. Tübingen: Universität Tübingen, 2010. <<http://www.tustep.uni-tuebingen.de/>>.

TUSTEP: Tübinger System von Textverarbeitungsprogrammen. Version 2010. Handbuch und Referenz. Tübingen: Universität Tübingen, Zentrum für Datenverarbeitung, 2009.

<<http://www.tustep.uni-tuebingen.de/pdf/handbuch.pdf>>.

Das MaGI-Projekt: Elektronische Katalogisierung der griechischen Handschriften Italiens

Marilena Maniaci, Paolo Eleuteri

Zusammenfassung

Größere und kleinere italienische Bibliotheken bewahren ca. 6500 griechische Handschriften. Dazu kommen ca. 4700 Bände der Vatikanischen Bibliothek. Diese Zahlen beruhen auf reinen Schätzungen, weil neuzeitliche Handschriften, in Archiven aufbewahrte Codices und insbesondere die Zahl der wirklich kodikologischen (also nicht nachträglich buchbinderischen) Einheiten unbekannt ist. Das griechische Handschriftenerbe Italiens muss in Umfang und Inhalt also noch genauer bestimmt werden, was jedoch durch das Fehlen angemessener Bestandsnachweise erschwert wird. Diesen Problemen will ein Langzeitprojekt abhelfen, das eine umfassende elektronische Katalogisierung und – zumindest teilweise – Digitalisierung aller griechischen Handschriften in italienischen Bibliotheken zum Ziel hat. Die Arbeit erfolgt dabei ›offen‹ und ›kollaborativ‹ mit Hilfe einer Software, auf der auch das Katalogisierungsprojekt der ›Nuova Biblioteca Manoscritta‹ beruht das seit 2003 die Katalogisierung der mittelalterlichen und neuzeitlichen Handschriften der Region Venezien betreibt. Diese Software kann einfach für die Katalogisierung von Handschriften unterschiedlicher kultureller Herkunft angepasst werden und ist mit dem vom Istituto Centrale per il Catalogo Unico gepflegten Standard ›Manus‹ kompatibel. Die Ergebnisse der Katalogisierung werden den Forschern frei zur Verfügung stehen und können für paläographische, kodikologische, kunsthistorische und philologische Forschungen an den beschriebenen Handschriften benutzt werden. Da vergleichbare nationale oder internationale Projekte zu byzantinischen Handschriften fehlen, wird ›MaGI‹ auch neue Referenzwerke zur Handschriftenbeschreibung erarbeiten und bereitstellen: Thesauri für Autorennamen und Buchtitel der griechischen Klassik und der byzantinischen Zeit. Diese werden sowohl für die Katalogisierung von Handschriften als auch von Drucken in diesem Feld dienen können.

Abstract

The number of Greek manuscripts in major and minor Italian libraries amounts to ca. 6,500 volumes (not counting ca. 4,700 units belonging to the Vatican Library). These figures however are only roughly approximate, because of the unspecified amount of modern manuscripts in libraries and archives and above all to the unknown

quantity of codicological units composing each volume. Extent and typology of the Greek manuscript heritage preserved in Italian libraries remains to be ascertained with an acceptable precision, and its proper understanding is made difficult by the lack of adequate research tools. This justifies the need for a long-term project, aiming, as a final task, at the electronic cataloguing and the (at least selective) digitisation of all the Greek codices held in Italian libraries. Cataloguers will work in an ›open‹ and ›collaborative‹ environment, using a specific software created on the basis of the existing »Nuova Biblioteca Manoscritta«, developed in 2003 for the description of the medieval and modern manuscripts of the Regione Veneto. This choice has been made considering not only the flexibility of the software (which can be easily adapted to manuscripts of different cultural areas), but also its full compatibility with the »Manus« standard, maintained by the Istituto Centrale per il Catalogo Unico. The results will be made freely available to the scientific community and will also be used as the basis for specific research concerning the palaeographical, codicological, art-historical and textual features of the described manuscripts. Due to the lack of related national and international projects in the field of the Byzantine book research, the project »MaGI« is also meant to provide new tools for cataloguing, such as a set of authority lists for classical and Byzantine names and titles, to serve as a tool for referencing both, the cataloging of manuscripts and printed books.

1. Einführung: Griechische Handschriften in Italien

Es ist eine bekannte Tatsache, daß Italien einen größeren Bestand an griechischen Handschriften besitzt als alle anderen westlichen Länder, sieht man einmal von der ›Ausnahme‹ Griechenland ab. Denn im Unterschied zum übrigen Europa sind in Italien (insbesondere in den mittel- und süditalienischen Regionen) das gesamte Mittelalter hindurch griechische Handschriften abgeschrieben worden. Dem Ausklingen des langen und lebhaften Zeitalters der italo-griechischen Buchkultur war das wiedererwachte Interesse an der griechischen Sprache im frühen Humanismus gefolgt. Dabei wurden Autoren und Codices wiederentdeckt sowie Hof- und Privatsammlungen geschaffen. Schließlich ist die Abwanderung von griechischen Gelehrten und Kopisten nach Italien schon vor, vor allem jedoch nach der Eroberung Konstantinopels durch die Türken zu nennen. Diese Migration war ihrerseits Ursache für eine neue Welle von Abschriften durch sowohl griechische als auch italienische Kopisten und zwar bis weit ins 16. Jahrhundert hinein. Schon seit Anfang des 15. Jahrhunderts bis in die Zeit der ausgehenden Renaissance haben namhafte Sammlungen (wie die Biblioteca Laurenziana in Florenz, die Biblioteca Marciana in Venedig, die Biblioteca Ambrosiana in Mailand, die Nationalbibliothek in Neapel oder natürlich auch die Vatikanische Bibliothek) weitreichend und systematisch Bestände aus dem griechischen Osten und aus den basilianischen Klöstern in Süditalien erworben; hinzukommen die üblichen

Hinterlassenschaften und Schenkungen. Schon in dieser Epoche zeichnet sich die Physiognomie eines Bibliothekenschatzes ab, der in den darauffolgenden Jahrhunderten weiter zugenommen hat.

Die Gesamtzahl der griechischen Handschriften in italienischen Bibliotheken wurde von Mioni auf etwas mehr als 6600 Stück geschätzt (etwa ein Fünftel der Gesamtzahl überhaupt),¹ einschließlich einer nicht genau zu bestimmenden Anzahl aus dem 16. Jahrhundert, einer Zeit, in der die Produktion von griechischen Handschriften nicht so schnell vom Buchdruck verdrängt wurde wie es im lateinischen Westen der Fall ist (cf. zuletzt Maniaci).

Ebenso wie die lateinischen sind auch die griechischen Handschriften Italiens dicht, aber quantitativ uneinheitlich verstreut – mit anderen Worten: 6164 Handschriften und damit fast die Gesamtzahl der Handschriften überhaupt befinden sich in den Bibliotheken von nur zehn Städten. Es handelt sich um die florentinischen Bibliotheken (angeführt von der Laurenziana), die Marciana in Venedig, die Ambrosiana in Mailand, die römische Trias (Angelica, Vallicelliana und Casanatense), das Zönonium von Grottaferrata, die Nationalbibliotheken von Neapel und Turin, die Biblioteca Estense in Modena, die Regional- und Universitätsbibliothek in Messina, sowie die beiden bologneser Bibliotheken, Universitätsbibliothek und Archiginnasio.

Die übrigen, etwas weniger als 500 griechischen Codices sind hingegen auf insgesamt 58 verschiedene Standorte und Institutionen verteilt (in kleinen und kleinsten Sammlungen von 1–50 Exemplaren): Bibliotheken (National-, Staats-, Provinz-, Kommunal-, Universitäts-, kirchliche und private Bibliotheken), Staats- und Kirchenarchive, Kirchengemeinden und Klöster, Privatsammlungen (von Laien und Geistlichen) – mit den entsprechend großen Schwierigkeiten, was Sachkenntnis, Verwaltung und Erhaltung angeht.

Dank der sachkundigen Bemühungen, die kontinuierlich vom 18. Jahrhundert bis auf den heutigen Tag andauern, kann der Forscher, der sich mit griechischen Codices Italiens unter philologischen Gesichtspunkten beschäftigt, zum großen Teil auf präzise, detaillierte Beschreibungen zurückgreifen, die als solide Grundlage für weitere Aktualisierungen und Verbesserungen nutzbar sind. Allerdings sind sie in manchen (und darunter wichtigen) Fällen sehr veraltet: Man denke nur an die über 1200 griechischen Codices der Laurenziana, für die noch immer die verdienstvollen, aber veralteten Beschreibungen von Angelo Maria Bandini dienen, oder an die nahezu 1100 Codices der Ambrosiana, die zwischen 1894 und 1906 beschrieben worden sind.

Die augenfälligsten Mängel der bis Mitte des letzten Jahrhunderts erstellten Kataloge betreffen verständlicherweise den Umgang mit äußeren Aspekten. Unvollständig – und zumeist fehlerhaft oder unzweckmäßig – sind zum Beispiel die Angaben über die

¹ Nach Hunger (43), der die Gesamtzahl der erhaltenen griechischen Handschriften auf etwa 30000 schätzt. Höher (bei etwa 47000 Exemplaren) liegt die Zahl, die sich Richard und Olivier zufolge ermitteln lässt, die aber eine nicht genauer bestimmbare Anzahl moderner Handschriften mit einschließt.

Schreibstoffe, die Liniierung oder die »mise en page«. Noch gravierender ist allerdings, daß nur die neueren Kataloge eine Rekonstruktion der Struktur der Handschriften erlauben, also des Bezugs der Handschriften zu ihrem Inhalt und zu den eventuellen Veränderungen dieses Verhältnisses. Trotz der in der griechischen Paläographie und Kodikologie in den letzten Jahrzehnten erreichten Fortschritte ist die Kenntnis der Herstellungstechniken griechischer Handschriften in verschiedener Hinsicht noch begrenzt; es fehlen genügend große und systematische Sammlungen brauchbarer Daten für quantitative Untersuchungen. Beispielhaft sind zu nennen: die genaue Charakterisierung und zeitliche und räumliche Verteilung der verschiedenen Arten von wasserzeichenlosem Papier, die in Byzanz verwendet wurden; die systematische Verzeichnung der in den Wasserzeichen verwendeten Motive; die Techniken und Werkzeuge, die für die Liniierung der Papierhandschriften verwendet wurden; die Kriterien der Herstellung des Blatts und des Füllens und Ausnutzens der Seite in spätbyzantinischen Handschriften; die Prinzipien, die die Zusammenfügung verschiedener Texte zu einer »kodikologischen Einheit« regeln und die Zusammenfügung bzw. Trennung, die im Laufe der Zeit aus einzelnen Einheiten einen Band bzw. aus einem Band mehrere Bände gemacht haben.

Die »älteren« Kataloge sind weiterhin hinsichtlich ihrer Angaben zur Datierung der Handschriften problematisch. Diese beruhen auf Schätzungskriterien, die seitdem immer wieder durch die Fortschritte der Paläographie grundlegend in Frage gestellt worden sind. Mangels einer hinreichenden fotografischen Dokumentation ist die Berichtigung der Fehler auf eine Sichtung der Codices selbst angewiesen und damit abhängig von gelegentlichen Resultaten einzelner Forschungsprojekte.

Die fotografische Reproduktionen der griechischen Handschriften Italiens ist aber – trotz des Vorhandenseins von paläographischen Alben, Sammlungen von datierten Codices, Repertorien von Kopisten – ziemlich spärlich, sowohl quantitativ als auch typologisch repetitiv sowie in Publikationen von unterschiedlicher Art und Zugänglichkeit zerstreut. Eine weiträumige Verfügbarkeit von Abbildungen für einen gezielten Zugriff würde nicht nur die Berichtigung der Datierungsfehler erleichtern, sondern sicher auch zur Fortentwicklung der bestehenden paläographischen Kenntnisse beitragen. Die Geschichte der Formen, Entwicklungen und Verwendungsweisen der griechischen Schrift ist in der Tat noch eher lückenhaft beschrieben oder enthält zumindest hoch problematische Teile, deren Erforschung unzweifelhaft von neuem bzw. kaum bekanntem Material stark profitieren könnte.

2. Das Projekt MaGI

Die Existenz einer beträchtlichen, aber dennoch überschaubaren Anzahl von in Italien verwahrten griechischen Codices hat eine Gruppe von Universitätsdozenten und

Konservatoren² veranlasst, ein Projekt für eine Sammeldatenbank dieser Handschriften zu entwickeln, die folgende Zielsetzungen verfolgt:

- eine zusammenfassende Online-Katalogisierung der griechischen Handschriften in Italien, organisiert in Form einer Datenbank, die sowohl für punktuelle als auch systematische paläographische und kodikologische Untersuchungen zur Verfügung steht;
- ein »digitales paläographisches Album« der griechischen Handschriften in Italien mit mindestens einer fotografischen Reproduktion für jede kodikologische Einheit (im Idealfall eine Reproduktion pro nachgewiesenem Schreiber, plus andere paläographische, kodikologische und dekorative Einzelheiten);
- eine bibliographische Datenbank, die in Zusammenarbeit mit den Autoren von »Pinakes« (dem Verzeichnis griechischer und byzantinischer Texte, herausgegeben von der »Section grecque« des IRHT), aufgebaut wird (cf. zuletzt Binggeli und Cassin).

Dieser zweckmäßigen Ausrichtung folgend wurde eine Initiative zur systematischen Erschließung der griechischen Handschriften in Italien unter dem Namen »MaGI« (Manoscritti Greci d'Italia) gestartet. Dieses Projekt kann die jüngsten Entwicklungen innerhalb namhafter Katalogisierungsvorhaben nicht außer Acht lassen; Entwicklungen, die immer mehr auf eine primäre und exklusive Nutzung des Internets hinauslaufen, auch wenn noch aussagekräftige Beispiele von im Internet verbreiteten Katalogen griechischer Handschriften ausstehen.³ Die elektronische Katalogisierung der griechischen Handschriften in Italien versteht sich aber nicht nur als Instrument für eine bessere wissenschaftliche Kenntnis dieser Handschriften, sondern setzt sich noch ein weiteres Ziel: Durch die Verwendung einer gemeinsamen Terminologie und den Gebrauch übereinstimmender Namen und einheitlicher Werktitel will diese Katalogisierung einen Beitrag zur Festlegung eines Standards der Handschriftenbeschreibung leisten. Von einem solchen Standard, anhand dessen sich griechische Handschriften umfassend beschreiben liessen, ist man noch weit entfernt. Bis heute gibt es keine maßgebliche Liste der griechischen klassischen und byzantinischen Namen; eine solche Liste ist bisher weder von der International Federation of Library Associations and Institutions (IFLA) noch vom Istituto Centrale per il Catalogo Unico (ICCU) vorgelegt worden. In den neuen italienischen Katalogisierungsregeln (REICAT 2009) ist für die klassischen und byzantinischen griechischen Namen lediglich festgelegt, daß die jeweils gebräuchliche lateinische bzw. latinisierte Form zu verwenden ist. Aber

² Zu der Gruppe gehören, neben Paolo Eleuteri und Marilena Maniaci, die Kolleginnen und Kollegen Edoardo Crisci, Paola Degni, Maria Rosa Formentin, Margherita Losacco, Pasquale Orsini und Elisabetta Sciarra.

³ So werden beispielsweise bei der Bestandsaufnahme der Handschriften der italienischen Bibliotheken durch das Istituto Centrale per il Catalogo Unico nur Handschriften in lateinischer Schrift berücksichtigt.

es ist offenkundig, daß es angesichts fehlender Referenzausgaben der Werke der klassischen und vor allem der byzantinischen Autoren schwierig sein wird, den Namen zu verwenden, der sich in den Werkausgaben am häufigsten findet, bzw. den, der aus bibliographischen Nachschlagewerken bezogen ist. Bereits die italienischen Regeln für die Katalogisierung von Autoren (RICA 1979) ignorieren den Fall byzantinischer Autoren, deren Werke in moderner Zeit nur auf griechisch ediert worden sind und von denen es keine allgemeingebräuchliche latinisierte Form gibt. Bis heute bestehen verschiedene Referenz-Repertorien: Volpi (dieses grundlegende Werk basiert auf anderen gebräuchlichen Werken wie etwa Geerard, *CPG*; Berkowitz, Squitier); ferner die autoritativen Verzeichnisse der Biblioteca di cultura medievale der Fondazione Ezio Franceschini und ebenso das Verzeichnis der Datenbank *Pinakes: Textes et manuscrits grecs*, herausgegeben vom Institut de recherche et d'histoire des textes (IRHT). Dieses letztgenannte Hilfsmittel ist zwar das für die byzantinischen Namen umfangreichste, zeigt jedoch zahlreiche Zweifelsfälle bei der Bestimmung von Autoren und der ihnen zugeschriebenen Werke.

Als Software für die Eingabe und Verwaltung der Beschreibungen wurde die Plattform »Nuova Biblioteca Manoscritta« gewählt. Die Kodierung des Griechischen erfolgte mittels eines Unicode-Fonts, und das Exportieren der Dateien erfolgte über das Manuscript Description Modul der TEI. Diese Nuova Biblioteca Manoscritta (NBM) ist im Prinzip ein OPAC der Handschriften der Bibliotheken des Veneto ohne zeitliche und inhaltliche Beschränkung. Mehr als 90000 Handschriften werden in Bibliotheken des Veneto insgesamt aufbewahrt, nicht eingerechnet Briefwechsel und andere handschriftliche Bestände, die nicht in Codexform aufbewahrt sind. Das Projekt, in seiner Form einzigartig in Italien und finanziert von der Landesregierung des Veneto in Zusammenarbeit mit der Universität Venedig, wurde 2003 begonnen und weist inzwischen Bestände aus 40 Bibliotheken nach. Dieser offene Katalog ermöglicht es anderen Bibliotheken, ihre Bestandsinformationen ebenfalls einzuspeisen. Je nach den eigenen spezifischen Anforderungen und Gegebenheiten kann die Katalogisierung sowohl auf einem einfachen Minimalniveau erfolgen als auch weitergehenden wissenschaftlichen Ansprüchen Genüge leisten – dies alles auf der Grundlage verschiedener kodikologischer Einheiten. Bis heute sind im Rahmen der NBM die Beschreibungen von mehr als 27000 Handschriften veröffentlicht (Bernardi et al.).

Die NBM ist Internet-basiert, das heißt, daß ihr Kern eine zentrale Datenbank ausmacht, in die teilnehmende Bibliotheken ihre Daten einspeisen. Die Bearbeiter nutzen zentrale, gemeinsame Thesauri für Namen, Titel, Schlagwörter, alte Signaturen, bibliographische Angaben, Textarten und Genres. Sie haben Zugriff auf die schon eingespeisten Daten, die sie im Sinne eines offenen Katalogs kontinuierlich aktualisieren. Der Katalog bietet die Möglichkeit, einzelnen Teilen der Beschreibung ebenso wie der gesamten Handschrift Bilder beizugeben. Ebenso lassen sich Bilddigitalisate ganzer Handschriften einfügen, die dann seitenweise durchgeblättert werden können. Die

NBM wird von der primären Datenerfassung über Revisionen bis zur abschließenden Veröffentlichung vollständig über das Internet verwaltet. Das System bietet Profile für verschiedene Benutzerrollen: Ein Koordinator richtet Benutzerkonten für die Katalogisierung ein, kontrolliert und begutachtet die Datensätze und überwacht den Inhalt des gesamten Webangebots. Richtlinien zur Katalogisierung stellen ein Maximum an Einheitlichkeit bei den Beschreibungen sicher. Die Datenbank kann über einen OPAC auf der Website und über das Z39.50-Protokoll nach verschiedenen Kriterien abgefragt werden.⁴

In jüngster Zeit ist eine neue Anwendung der NBM hinzugekommen. Es ist jetzt möglich, nicht mehr nur auf der Basis von Projekten zu arbeiten, die an eine Bibliothek gebunden sind, sondern man kann nun eine Datenbank für bibliotheksübergreifende Projekte nutzbar machen. Die Listen von Namen und Überschriften, die Bibliographie, Inhaltsangaben und Angaben zu Bibliotheken lassen sich für verschiedene Bibliotheken und Gruppen von Wissenschaftlern zugänglich machen, die nicht einmal unbedingt mit demselben Alphabet arbeiten.

Von der MaGI Projekt-Homepage aus hat man Zugang zu allgemeinen Informationen über das Projekt sowie den Katalog, den teilnehmenden Bibliotheken, einer digitalen Bibliothek, einer Sammlung von Links, den Richtlinien für die Katalogisierung und den Adressen der Mitarbeiter. Dank der Finanzierungszusagen der Universitäten von Cassino und Venedig war es möglich, die Katalogisierung und Dokumentierung einiger Handschriftenfonds auf den Weg zu bringen: Dazu gehören das Archiv der Abtei von Montecassino (zusammen mit den entsprechend korrigierten Angaben aus dem neuen Katalog von Patrizia Danella), ferner die römischen Bibliotheken⁵ und schließlich, seit dem Beginn der Erneuerung des Katalogs der griechischen Handschriften, die Marciana in Venedig.

Das MaGI Projekt ist nicht nur ein Werkzeug zur umfassenden Erschließung und Bewertung des griechischen Handschriftenerbes in Italien. Angestrebt ist darüber hinaus die Schaffung eines Prototypen für vergleichbare internationale Initiativen für griechische Handschriften, die im Internet erst spärlich vertreten sind. MaGI soll ein Modell für eine mögliche Nutzbarmachung von Hilfsmitteln und Ergebnissen sein. Als ein solches Modell steht MaGI auch in den vor dem Abschluß stehenden Vereinbarungen mit dem Schweizerischen Nationalfonds über die Zusammenarbeit bei der wissenschaftlichen Erforschung und Katalogisierung der griechischen Handschriften in schweizerischen Bibliotheken und ebenso in den Vereinbarungen mit der Section

⁴ Eine Beschreibung der Funktionalität der NBM kann man in den verschiedenen Publikationen von Eleuteri, Vanin und Bernardi finden, die in der Bibliographie aufgelistet sind.

⁵ Für die Angelica ist der Abbildungsband dank der Arbeit von Elisabetta Sciarra fertiggestellt und seinerseits mit dem gescannten Katalog aus dem 19. Jahrhundert von Giorgio Muccio e Pio Franchi de' Cavalieri versehen; die Bestände der Casanatense und der Vallicelliana werden von Pasquale Orsini katalogisiert.

grecque des IRHT über die Herausgabe einer Bibliographie der griechischen Codices, die sowohl in MaGI als auch in Pinakes zu finden sind.

Bibliographie

- Berkowitz, Luci und Karl A. Squitier. *Canon of Greek Authors and Works*. New York – Oxford: Oxford University Press, 1990.
- Bernardi, Francesco, Paolo Eleuteri und Barbara Vanin. »La catalogazione in rete dei manoscritti delle biblioteche venete: Nuova Biblioteca Manoscritta.« *KPDZ* 1. 3–11.
- Binggeli, André und Mathieu Cassin. »Recenser la tradition manuscrite des textes grecs: du Greek Index Project à Pinakes.« *La descrizione dei manoscritti: esperienze a confronto. Atti dei seminari di Cassino, 15 aprile 2008 e 18 novembre 2009*. Hg. Edoardo Crisci, Marilena Maniaci und Pasquale Orsini. Cassino: Dipartimento di Filologia e storia, 2010. 91-106.
- Bandini, Angelo Maria. *Catalogus codicum manuscriptorum Bibliothecae Mediceae Laurentianae varia continens Opera Graecorum Patrum sub auspiciis Francisci imp. semper augusti Ang. Mar. Bandinius i.v.d. eiusdem bibliothecae regius praefectus recensuit, illustravit, edidit. In eo cujusvis codicis accurata descriptio & operum singulorum notitia datur, vetustiorum specimina exhibentur, edita supplentur & emendantur. Plura adcedunt anecdota, pleraque Latine reddita*. Vol. I: Florentiae: Typis Caesareis, 1764; Vol. II: Florentiae: Typis Regiis, 1768; Vol. III: Florentiae: Typis Regiis, 1770.
- Clavis Patrum Graecorum*. Hg. Maurice Geerard et al. Turnhout: Brepols, 1974–1998.
- Eleuteri, Paolo. »La catalogazione in rete dei manoscritti delle biblioteche venete.« *Zenit e Nadir II. I manoscritti dell'area del Mediterraneo: la catalogazione come base della ricerca. Atti del seminario internazionale (Montepulciano, 6–8 luglio 2007)*. Hg. Benedetta Cenni, Chiara Maria Francesca Lalli und Leonardo Magionami. Montepulciano: Thesan e Turan, 2007. 221–225.
- Eleuteri, Paolo und Barbara Vanin. »Il catalogo on line dei manoscritti delle biblioteche del Veneto.« *Gazette du livre médiéval* 47 (2005): 31–38.
- Eleuteri, Paolo und Barbara Vanin. »Nuova Biblioteca Manoscritta«. Catalogo dei manoscritti promosso dalla Regione del Veneto.« *La descrizione dei manoscritti: esperienze a confronto. Atti dei seminari di Cassino, 15 aprile 2008 e 18 novembre 2009*. Hg. Edoardo Crisci, Marilena Maniaci und Pasquale Orsini. Cassino: Dipartimento di Filologia e storia, 2010. 61-69.
- Hunger, Herbert. *Schreiben und Lesen in Byzanz. Die byzantinische Buchkultur*. München: C. H. Beck, 1989.
- MaGI. Manoscritti greci d'Italia*. [Venezia: Nuova Biblioteca Manoscritta], 2010.
<<http://www.nuovabibliotecamanoscritta.it/MaGI/>>.
- KPDZ 1: *Kodikologie und Paläographie im Digitalen Zeitalter / Codicology and Palaeography in the Digital Age*. Hg. Malte Rehbein, Patrick Sahle und Torsten Schaßan. Norderstedt: Books onDemand, 2009. Online: <urn:nbn:de:hbz:38-29393>,
<<http://kups.ub.uni-koeln.de/volltexte/2009/2939/>>.
- Martini, Emilio und Domenico Bassi. *Catalogus codicum graecorum Bibliothecae Ambrosianae*, I–II. Milano: Hoepli, 1906.

- Maniaci, Marilena. »La catalogazione dei manoscritti greci conservati in Italia: stato e prospettive.« *La descrizione dei manoscritti: esperienze a confronto. Atti dei seminari di Cassino, 15 aprile 2008 e 18 novembre 2009*. Hg. Edoardo Crisci, Marilena Maniaci und Pasquale Orsini. Cassino: Dipartimento di Filologia e storia, 2010. 107–127.
- Mioni, Elpidio. »La catalogazione dei manoscritti greci in Italia.« *Il manoscritto. Situazione catalografica e proposta di una organizzazione della documentazione e delle informazioni. Atti del seminario di Roma (11–12 giugno 1980)*. Hg. Maria Cecilia Cuturi. Roma: ICCU, 1981. 15–25.
- Nuova Biblioteca Manoscritta (NBM). *Catalogo dei manoscritti delle biblioteche del Veneto*. [Venezia, 2006–2010] <<http://www.nuovabibliotecamanoscritta.it/>>.
- Pinakes. *Textes et manuscrits grecs*. Paris: Institut de recherche et d'histoire des textes, 2008–2010. <<http://pinakes.irht.cnrs.fr/>>.
- REICAT. *Regole italiane di catalogazione*. Roma: ICCU, 2009.
- RICA. *Regole italiane di catalogazione per autori*. Roma: ICCU, 1979 (mehrfach nachgedruckt).
- Richard, Marcel und Jean-Marie Olivier. *Répertoire des bibliothèques et des manuscrits grecs de Marcel Richard*. Troisième éd. entièrement refondue. Turnhout: Brepols, 1995.
- Vanin, Barbara und Paolo Eleuteri. »Nuova Biblioteca Manoscritta. Catalogo in linea dei manoscritti delle biblioteche del Veneto.« *Bollettino dei Musei Civici Veneziani* 3a seria, 1 (2006): 113–117.
- Vanin, Barbara und Paolo Eleuteri. »La Nuova Biblioteca Manoscritta della Regione del Veneto.« *Conoscere il manoscritto: esperienze, progetti, problemi. Dieci anni del progetto Codex in Toscana*. Hg. Michaelangiola Marchiaro und Stefano Zamponi. Firenze: SISMEL – Edizioni del Galluzzo, 2007. 145–152.
- Vanin, Barbara. »»Nuova Biblioteca Manoscritta« Online Catalogue of Manuscripts Conserved in Libraries in the Veneto Region«. *Encyclopedia of Information Communication Technology (ICT)*. Hg. Antonio Cartelli und Marco Palma. Hershey (PA): Information Science Reference, 2008. 632–634.
- Volpi, Vittorio. *DOC. Dizionario delle opere classiche*. Milano: Editrice Bibliografica, 1994.

La numérisation du patrimoine livresque médiéval : avancée décisive ou miroir aux alouettes ?

Ezio Ornato

Résumé

La numérisation de la totalité du patrimoine livresque médiéval, qui pouvait paraître une utopie il y a quelques années, tend à devenir une perspective raisonnable dans un avenir relativement proche. Bien que de telles opérations doivent être nécessairement promues et gérées par les bibliothèques elles-mêmes, on peut se demander jusqu'à quel point les modalités de la numérisation seront compatibles avec les exigences des chercheurs : on peut craindre, en effet, un certain nombre de distorsions en ce qui concerne le choix des livres à numériser, la définition et la disponibilité effective des images et l'accès direct aux livres qui auraient été numérisés. Cela dit, la numérisation n'est pas une fin en soi. Elle devrait être considérée comme le moment clé d'une initiative plus vaste, ayant pour objectif la constitution d'une véritable *Bibliotheca universalis*. Or, cette bibliothèque virtuelle ne saurait fonctionner de manière satisfaisante en l'absence d'un arrière-plan descriptif adéquat, ce qui soulève beaucoup de problèmes. Cette constatation renvoie tout naturellement à l'interrelation entre la numérisation de l'objet et sa description catalographique qui, lorsqu'il s'agit de définir des priorités, pourrait donner lieu des choix difficiles. Cet antagonisme potentiel pourrait être dépassé grâce à la mise en œuvre sur le web d'un nouveau type de catalogue interactif qu'on pourrait appeler « grand-ouvert ». Il ne faut pas oublier, cependant, que le concept de *Bibliotheca universalis* présuppose que l'on puisse naviguer à loisir dans l'ensemble des images numérisées, et donc que les bases descriptives sous-jacentes soient interconnectées. Pour que cela soit possible, il faudrait que les données fondamentales soient harmonisées grâce à l'existence d'une métabase d'« autorités ». Nous en sommes loin. Il n'est pas sûr, enfin, que le chercheur puisse tirer pleinement parti de l'ensemble des données textuelles et iconographiques dans la perspective d'un véritable traitement de l'information : en effet, la structure et le contenu des bases ne sont pas assez rigoureux et les buts poursuivis se limitent le plus souvent à faciliter le repérage et la sélection d'objets possédant une ou plusieurs caractéristiques communes. A tout cela, il convient d'ajouter des empêchements qui découlent, malheureusement, de points de vue moins scientifiques qu'administratifs, voire lucratifs, chez les institutions.

Zusammenfassung

Die Digitalisierung sämtlicher mittelalterlicher Buchbestände, vor einigen Jahren noch utopisch, nimmt mehr und mehr die Gestalt einer realistischen Zukunftsperspektive an. Ein solches Vorhaben müsste notwendigerweise von den Bibliotheken selbst getragen und vorangetrieben werden. Es stellt sich dann aber die Frage, wie weit sich die Modalitäten einer solchen Digitalisierung auf die Bedürfnisse der Forschung zuschneiden ließen. Dabei könnte es zu einer Reihe von Zielkonflikten kommen: bei der Auswahl der zu digitalisierenden Bücher, der Auflösung und der Verfügbarkeit der Bilder sowie bei der Frage, welchen Benutzungseinschränkungen die Originale, die bereits als Digitalisat vorliegen, unterworfen werden. Davon abgesehen ist Bild-Digitalisierung kein Selbstzweck. Sie sollte vielmehr den entscheidenden Schritt einer sehr viel weiterreichenden Initiative darstellen, die sich die Errichtung einer *Bibliotheca universalis* zum Ziel setzt. Eine solche virtuelle Bibliothek würde allerdings nicht zufriedenstellend funktionieren, solange das zugrunde liegende Konzept nicht in angemessener Weise modelliert ist. Und hier fangen die Probleme an: Gerade im Verhältnis von Digitalisierung und Katalogisierung wären schwierige Entscheidungen für eine Prioritätensetzung zu fällen. Eine mögliche Lösung wäre dadurch zu erreichen, dass man im Internet einen neuartigen, interaktiven Katalog ins Leben rief, der sich als „grand-ouvert“, als „weit offen“ bezeichnen ließe. Dabei darf nicht außer Acht gelassen werden, dass das Konzept der *Bibliotheca universalis* voraussetzt, dass der Benutzer unbeeinträchtigt zwischen allen Digitalisaten navigieren kann und die zugrunde liegenden Beschreibungen sinnvoll verknüpft sind. Um das zu erreichen, müssten die Daten über einen noch zu schaffenden Thesaurus standardisiert werden, wovon wir jedoch noch weit entfernt sind. Es bleibt ungewiss, ob Forscher vollen Nutzen aus den angesammelten Text- und Bilddaten im Sinne einer wirklichen Informationsverarbeitung werden ziehen können, denn Struktur und Inhalt der Datenbanken sind nicht klar genug, und Zielvorgaben beschränken sich meist auf Kennzeichnung und Auswahl von Objekten mit einem oder mehreren einfachen Merkmalen. Hinzu kommen für gewöhnlich noch Zugriffsbeschränkungen, die leider weniger aus wissenschaftlichen denn aus ökonomischen Interessen der bewahrenden Institutionen erlassen werden.

Abstract

The complete digitisation of all collections of medieval books has long been regarded as an utopian goal. Now, however, this ambition is developing more and more into a realistic perspective for the future. Of course, such an enterprise needs to be initiated and driven by the libraries themselves. But it has to be clarified to what extent research and its needs and requirements have to be taken into account in modeling

this digitisation effort. A couple of conflicts of interest need to be considered: selecting the objects for digitisation, resolution and accessibility of the images, usage of the originals after digitisation etc. Furthermore, mere digitisation is not an end in itself but should become the core of a far-reaching initiative aimed at the establishment of a *bibliotheca universalis*. Such a virtual library would, however, be insufficient without a well designed underlying data model. And this is where the story gets really challenging: the relationship between digitisation on the one hand and cataloging on the other forces us to make severe decisions. A possible solution for this area of tension might be the establishment of a new kind of interactive catalogue in the internet: the “catalogue grand-ouvert”. Yet, one should not forget that the model of the *bibliotheca universalis* allows its user to freely navigate between all available digital objects and that this requires meaningful linkages among the descriptive data. To accomplish this, all data would need to be standardised by means of an ontology that has yet to be defined, and we are still far away from this.

1. Numériser l'intégralité du patrimoine manuscrit médiéval : est-ce faisable ?

Dans un article rédigé il y a quelques années, je préconisais la création d'une *Bibliotheca manuscripta universalis* dont le point fort serait la numérisation et la mise à disposition sur le Web de la totalité du patrimoine manuscrit conservé dans nos bibliothèques.

Bien sûr, je me suis vite aperçu que le terme *manuscripta* était à la fois trop large — car il s'agissait essentiellement de la production médiévale — et trop réducteur, car il excluait la production imprimée : l'expression *Bibliotheca universalis librorum Medii Aevi* aurait exprimé le concept de manière plus pertinente. Toutefois, malgré quelques remarques plutôt sarcastiques sur la faisabilité d'une telle entreprise, et bien que le mot « utopie » figurât prudemment dans le titre, je dois avouer que ce vœu me paraissait empreint de bon sens et somme toute relativement modeste dans ses ambitions. Cette étrange manifestation d'optimisme était, à mes yeux, justifiée par une série de considérations.

En premier lieu, les performances des systèmes informatiques, sur le plan aussi bien matériel que logiciel, ne cessent de progresser. Ces systèmes sont aujourd'hui capables de mémoriser et de traiter efficacement en un temps très court une immense quantité d'information : Google, Facebook, Youtube et bien d'autres initiatives, quelles que soient les réticences que l'on peut manifester à leur égard, nous font la démonstration quotidienne de ce que la technologie est déjà mille lieues en avance sur toute initiative que le médiéviste le plus hardi et le plus mégalomane serait capable de caresser ; et même, disons-le, sur toutes les initiatives auxquelles il n'a pas encore songé... et ne songera peut-être jamais. Aussi, ne s'agit-il plus d'appliquer ironiquement la célèbre devise des étudiants parisiens en révolte : « Soyons réalistes, demandons l'impossible »,

mais plutôt de mettre en œuvre une démarche mentale que dans notre pauvre (au sens financier du terme) univers des « humanités » nous sommes trop habitués à refouler : « Soyons irréalistes, imaginons le possible ». N'oublions pas que le corpus de livres dont Google Books — le vrai, pas la pâle caricature qui nous fait miroiter l'appât de quelques pages d'un livre au milieu d'un océan de trous et de liens vers les sites marchands — envisageait la numérisation est sans commune mesure avec le nombre, pourtant élevé, d'éditions incunables et de manuscrits du Moyen Âge occidental que conservent nos bibliothèques.

Ceux qui ont fréquenté autrefois assidument les anciens « palais » de l'informatique — où la « salle machines », comme dans les anciens paquebots, était interdite au personnel non autorisé — peuvent seuls se rendre compte de la transformation radicale, à tous points de vue, que le rapport « homme/ordinateur » a subie en l'espace de quarante ans. A l'époque, c'était le chercheur, surtout en sciences humaines, qui devait se plier à des exigences draconiennes auxquelles il était parfaitement étranger. Faire imprimer un titre en Fortran, quelle prouesse ! Lancer un calcul en Cobol, quelle performance ! Codifier strictement toute information pour gagner quelques kilooctets de mémoire, demander au moyen d'une carte perforée que l'on monte notre bande magnétique, attendre fiévreusement pendant des heures le retour peu glorieux des listings parsemés d'erreurs de programmation, quelle apothéose de la frustration ! Aujourd'hui, au contraire, c'est l'informatique qui vient au-devant de nos moindres caprices... Mais prenons garde : *Timeo Danaos, et dona ferentes*.

En deuxième lieu, la supériorité des technologies numériques par rapport aux dispositifs analogiques est telle, tant en termes de résultats qu'en termes de coûts, qu'elle pourrait être comparée à celle de l'imprimé par rapport au manuscrit, d'autant plus que le passage du manuscrit à l'imprimé sanctionnait le triomphe du noir et blanc, alors que le passage du microfilm à l'image numérisée coïncide précisément avec une réappropriation de la couleur. Or, qui peut imaginer ne serait-ce qu'un seul instant que le manuscrit aurait pu gagner la bataille ? Ou que le 78 tours aurait pu survivre au microsillon ? Et le microsillon au CD-ROM ? Et le CD-ROM au téléchargement ou au streaming ? Et si le téléphone fixe coexiste encore avec son homologue portable, c'est uniquement parce que les tarifs de ce dernier sont maintenus à un niveau artificiellement élevé grâce à des procédés plus ou moins occultes et, surtout, plus ou moins justifiés.

En troisième lieu, je présentais que l'idée de numériser les fonds manuscrits ne tarderait pas à faire tâche d'huile, une fois qu'un certain nombre d'obstacles seraient levés et que les modalités de l'opération évitent soigneusement de porter atteinte aux intérêts et aux privilèges de qui que ce soit. D'une part, parce qu'une initiative telle que Google Books, encensée par les usagers virtuels et dénoncée en chœur par les ayants droit et les tenants de l'identité culturelle — et pour ces raisons largement médiatisée — allait montrer qu'un projet de numérisation globale pouvait faire partie de la réalité concrète et fonctionner à la perfection ; de l'autre, parce que, dans une

société se proclamant soucieuse de la préservation de son environnement et de plus en plus conditionnée par le respect du principe de précaution, toute initiative ressemblant de près ou de loin à une sauvegarde trouverait assez vite un terrain propice à son développement.

Lorsque quelques grammes de CO₂ en moins deviennent un argument de vente — certes « bidon », mais ô combien révélateur — pour les industriels de l'automobile, on se dit que les temps sont mûrs pour que même la numérisation des livres anciens rentre de plein droit dans l'aéropage du « politiquement correct ». Ainsi, ce qui pouvait être au départ le rêve de quelques esprits visionnaires s'est concrétisé peu à peu en un petit nombre d'initiatives d'avant-garde, puis dans des projets de grande envergure, et bientôt les grandes bibliothèques qui n'auront pas mis en chantier leur programme de numérisation seront taxées de passéistes ; car les images des manuscrits sont capables de livrer aux yeux de tout un chacun le spectacle « magique » d'une partie de notre passé, et puisque l'existence d'un public potentiellement réceptif à un spectacle quelconque suscite nécessairement l'intérêt des pouvoirs de toute espèce, l'argent qui va avec ne se fera pas attendre.

Enfin, il fallait tout de même considérer l'éventualité que les bibliothèques — qui devraient être tout naturellement le centre moteur de la numérisation — rechignent à s'aventurer dans un chemin qui leur paraîtrait trop onéreux en termes de finances et de ressources humaines. Pour venir à bout de cette difficulté, il suffisait que :

- a) La numérisation soit financée par un budget spécifique sans préjudice pour la dotation courante de l'institution ;
- b) La bibliothèque conserve la gestion intégrale et exclusive des images ;
- c) L'opération soit entièrement externalisée grâce à des appels d'offre prévoyant le respect d'un cahier des charges précis et rigoureux ;
- d) La numérisation soit clairement découplée de toute initiative de catalogage.

Puisque ces conditions tendent désormais à être satisfaites, la numérisation déjà bien avancée du fonds « Plutei » de la Bibliothèque Laurentienne dans le cadre du projet « Teca digitale » et le lancement préliminaire d'un projet analogue concernant à terme les 80000 manuscrits de la Bibliothèque vaticane laissent présager que « les jeux sont faits » : dès lors que l'Église elle-même vient inopinément se placer dans un créneau d'avant-garde, il n'y a vraiment plus rien à craindre.¹

¹ Et pourtant, ce n'est pas la première fois ! Dès 1991, en effet, la Bibliothèque vaticane avait réalisé un vidéo-disque iconographique (Baryla ; Baschet).

2. La numérisation, les bibliothèques et les aspirations des chercheurs

Cependant, le concept de *Bibliotheca universalis librorum Medii Aevi* ne se réduit pas à la numérisation des livres anciens d'une, de plusieurs ou même de chaque bibliothèque de conservation. Il englobe bien sûr l'accès virtuel à toute la production livresque conservée du Moyen Âge occidental (et d'autres aires géographiques si l'on veut), mais requiert aussi :

1. La création de bases de données pourvues de multiples fonctions, à savoir :
 - a) Fournir des informations aptes à guider préalablement la navigation à travers l'ensemble des livres numérisés.
 - b) Fournir les informations sur les caractéristiques codicologiques et textuelles des livres que l'on ne peut extraire à partir d'une reproduction.
 - c) Fournir, autant que possible, les informations qui, du fait qu'elles requièrent des compétences spécifiques, sont habituellement le fait des spécialistes des diverses disciplines : localisation, datation et provenance des livres, identification des copistes et des ateliers d'enluminure ; identification des auteurs et des textes.
 - d) Permettre la sélection de livres, ou de groupes de livres, qui possèdent en commun une ou plusieurs propriétés.
 - e) Pouvoir s'intégrer, le cas échéant, dans un réseau de bases de données relatives à l'histoire de la culture écrite.
 - f) Fournir des résultats structurés, permettant à tout utilisateur de construire par ses propres soins de nouvelles bases dans la perspective d'une recherche ciblée.
2. Le libre accès à toute la littérature scientifique concernant l'histoire de la culture écrite.

Le fait que les programmes en cours ou en projet mettent systématiquement l'accent sur la numérisation, et que les autres points soient tout aussi systématiquement ignorés, n'est pas anodin : cela signifie que le point de vue des institutions qui ont la charge de conserver le patrimoine est plus restreint que celui des chercheurs qui voudraient pouvoir y accéder à leur gré ; cela signifie, aussi, que les chercheurs concernés ne forment pas une communauté scientifique suffisamment active et consensuelle pour pouvoir formuler et justifier de manière claire et organisée un certain nombre d'exigences, et encore moins pour donner le jour à des initiatives concertées.

Pour les bibliothèques, en effet, l'objectif premier de la numérisation du patrimoine n'est pas celui d'en faire un outil performant et irremplaçable pour l'historien de la culture écrite. En fait, l'opération est porteuse à la base de trois connotations différentes : elle est à la fois un service rendu aux « usagers », une vitrine destinée au « public » et un dispositif de régulation dans le contexte de la conservation.

En tant que service rendu aux « usagers », elle remplace avantageusement le microfilm destiné autrefois aux lecteurs potentiels qui ne pouvaient se rendre sur place ou y séjourner pendant un laps de temps suffisamment long. L'avantage réside dans le fait qu'une image en couleurs immédiatement accessible, pouvant être multipliée à loisir et à moindre coût à partir d'un original, surclasse sur tous les plans l'image analogique en noir et blanc dont la reproduction demande du temps, du personnel et... de l'argent.

En tant que vitrine destinée à un « public », la numérisation s'inscrit à merveille dans la nouvelle philosophie qui a été imposée aux bibliothèques depuis une trentaine d'années et qui règne en souveraine à l'heure actuelle : l'institution doit justifier auprès du grand public l'argent que celui-ci dépense en tant que contribuable ; dans cette perspective, elle doit produire constamment des « événements », et donc exposer le plus possible ses « trésors ». Inversement, elle doit empêcher la dégradation du patrimoine dont elle a la charge ; patrimoine qu'il est donc commode de maintenir à l'abri dans d'obscurs magasins. Problème : entre la vitrine et les magasins, il y a tout de même une salle de « médiation » ; celle qui devrait mettre les témoins du passé au contact de ceux qui voudraient en prendre connaissance et qui ne devraient pas pour autant être considérés comme un « facteur de dégradation » supplémentaire. En substance, est-ce trop prétendre que le rôle d'une bibliothèque de conservation ne soit pas confondu avec celui d'un musée ?

Dans le domaine de la conservation, la numérisation joue un double rôle : direct d'une part, dans la mesure où, si l'objet matériel venait à disparaître suite à un événement catastrophique d'origine naturelle ou humaine, l'image numérisée en garderait malgré tout un souvenir beaucoup plus fidèle que ne pouvait le faire le pâle simulacre microfilmé ; indirect d'autre part, dans la mesure où la qualité de l'ersatz pourrait rendre plus acceptable le refus de l'accès à l'original.

En d'autres termes, la numérisation, vue par les bibliothèques — ou du moins par certaines d'entre elles — apparaît comme le point d'arrivée d'une opération stratégique, alors que, dans la perspective d'une *Bibliotheca universalis*, elle représente le point de départ indispensable d'une ouverture sans précédent. Dès lors, cette opération risque d'être envisagée comme un aspect de la *gestion* du patrimoine plutôt que comme l'une des étapes essentielles de sa *fructification* dans l'intérêt de la recherche historique.

Dans ces conditions, il y a lieu de craindre que les modalités de l'opération puissent être en quelque sorte « détournées » en fonction de finalités qui dans le meilleur des cas seraient étrangères aux aspirations des chercheurs et, dans le pire, incompatibles avec elles.

Un certain nombre de distorsions porte sur les principes inspirateurs du programme de numérisation et les choix qui en découlent. Le résultat est toujours une perte d'exhaustivité, ce qui est l'antithèse même du concept de *Bibliotheca universalis*.

La distorsion la plus répandue — et sans doute la plus difficile à éradiquer — pourrait être caractérisée, pour utiliser une expression imagée, par la célèbre répartie

« *Haec sunt ornamenta mea* ».² Elle consiste à restreindre la numérisation aux seuls manuscrits connus ; en général, ceux qui brillent par la richesse de l'apparat décoratif et illustratif. Cette solution représente le point de rencontre spontané entre le concept de « bibliothèque vitrine », l'engouement très compréhensible des bibliophiles pour les « beaux livres » et la tendance — beaucoup moins compréhensible — des historiens de l'art à ne regarder dans les livres que l'habillage iconographique.

Cette option, déjà peu enthousiasmante en tant que principe général, présente à son tour des variantes qui le sont encore moins : la numérisation des seules pages enluminées, la numérisation de quelques-unes parmi les pages enluminées ou pire, la numérisation des seules parties enluminées de quelques pages.

Il ne faut pas négliger non plus la distorsion de type identitaire qui consiste à restreindre la numérisation aux livres « autochtones » : éditions originaires d'un pays ou d'une ville, exemplaires uniques, manuscrits dont les textes sont écrits dans la langue nationale.

Citons également une solution bancale à tous points de vue : la numérisation des microformes en lieu et place des originaux ; opération qui devient encore plus incompréhensible lorsqu'on poursuit en même temps le microfilmage systématique des fonds au lieu d'en entreprendre la numérisation.

Il convient malgré tout de rappeler que ces distorsions ne relèvent pas toujours d'une position de principe : il s'agit souvent d'un compromis imposé par une dotation financière insuffisante. Dans ce cas, il faut nécessairement établir des priorités, dont le caractère provisoire risque malheureusement de s'éterniser.

D'autres distorsions, tout aussi redoutables, concernent non pas la numérisation elle-même, mais la gestion des images numérisées. Ainsi, ce dont on pourrait à première vue craindre la généralisation, c'est un accès aux images entièrement payant et, par conséquent, strictement réglementé.

Heureusement, il ne semble pas, désormais, que l'on s'achemine vers ce genre de solution. Jusqu'à il y a peu de temps, la politique ultralibérale menée par quelques pays dans le domaine de la culture avait contraint les bibliothèques à rentabiliser au maximum les services rendus. Mais aujourd'hui, personne n'accepterait de payer pour avoir accès sur le Web à des banques de données ou d'images dont la création et la mise en œuvre auraient été financées par de l'argent public. Aussi, n'ai-je trouvé sur le Web qu'un seul exemple de ce type de gestion : le projet « Innerio », relatif aux manuscrits du Collège d'Espagne à Bologne, dont le mode d'utilisation, de plus, est particulièrement contraignant : un contrat d'abonnement, en effet, n'est valable que pour un seul numéro d'IP.³

² « Voici mes bijoux ! ». Phrase prononcée, selon la tradition, par Cornelia, la mère des Gracques.

³ Dans des domaines voisins, nous pourrions citer des bases de données à vocation culturelle dont la consultation est malheureusement payante, comme celles qui sont regroupées dans des portails tels que

Une distorsion plus subtile, mais qui pourrait se révéler beaucoup plus dangereuse, pourrait bien nous rappeler une fable fort connue : « Le renard et la cigogne ».⁴ Il s'agit d'un type de gestion où l'accès libre et gratuit serait réservé à des images de qualité délibérément inférieure, insuffisante pour une utilisation scientifique, l'accès au matériel exploitable étant soumis à conditions : accès payant, accès *in loco*. Les motivations réelles de cette restriction ne sont pas toujours explicitées mais, en tout cas, ne semblent pas relever exclusivement de la sphère financière.

Enfin, une dernière distorsion est étroitement liée à la problématique de la conservation : le libre accès aux images sur le Web pourrait impliquer, en contrepartie, que l'accès aux originaux soit réduit au minimum et, à la limite, totalement interdit, du moins pour certains manuscrits. Cette possibilité est explicitement mentionnée par les gestionnaires du projet de numérisation de la Bibliothèque Laurentienne : « L'elevata risoluzione delle immagini master consentirà di limitare, laddove necessario, il ricorso diretto agli originali (favorendone così la tutela e la conservazione) [c'est moi qui souligne] » (Degl'Innocenti 109).

Cet aspect de la question est absolument crucial, surtout si la politique de restriction est appliquée à large échelle. Peut-on considérer que des livres faits pour défier les siècles — et dont les pertes sont moins imputables à l'usure matérielle ou à un quelconque processus d'interaction avec le milieu ambiant qu'à leur obsolescence à la fois fonctionnelle, culturelle et idéologique qui les a envoyés au rebut — doivent être gérés comme la grotte de Lascaux ? Personnellement — mais je ne suis pas le seul — je ne vois aucune différence, sur le plan cognitif, entre un livre détruit et un livre enfermé à jamais dans un coffre-fort.

Cette contradiction entre les exigences de la conservation et les aspirations de la recherche n'est pas nouvelle : « Dans le domaine qui nous concerne, celui du manuscrit, on voit se développer dangereusement une tendance qui s'apparente de près à cette proposition [celle d'enfermer définitivement les œuvres d'art dans un coffre-fort qui, à l'époque, avait été assimilée à une simple boutade] : on voudrait que les chercheurs se contentent de consulter les microfilms et laissent tranquilles les précieux originaux » (Rizzo 14). Ces quelques lignes datent du printemps ... 1984 (précision nécessaire : il s'agit d'une vraie date, et pas du titre du roman homonyme), lorsqu'une historienne de la culture humaniste s'était vue refuser l'accès à un manuscrit, d'abord parce qu'il fallait

Brepolis ou *Mirabile*. Leur gestion n'est toutefois pas assurée par des organismes publics, mais par des entreprises commerciales ou par des institutions privées.

⁴ Le renard invite la cigogne à partager son repas, mais celui-ci est servi dans des assiettes plates sur lesquelles le long bec de l'oiseau n'a aucune prise. La cigogne invite le renard à son tour, et ce dernier se révèle incapable d'introduire son museau dans l'étroite ouverture du bocal qui contient le repas.

préserver l'intégrité du précieux objet, puis, une fois arrachée l'autorisation, parce que le volume était exposé dans une salle à côté!⁵

C'est par ailleurs en 1989 que les rédacteurs de la *Gazette du livre médiéval* proposaient, certes sur un mode plaisant, une « Déclaration des droits du manuscrit, du lecteur et du conservateur » dont l'article premier était ainsi rédigé : « Le manuscrit est fait pour le lecteur.⁶ Le traitement qui lui est réservé ne peut avoir pour but que d'en assurer le libre accès à tous ». ⁷

Il faut également souligner que la « pression » qu'exerce le lecteur sur l'intégrité du patrimoine manuscrit est largement surestimée. Sur ce point, nous ne possédons qu'une seule statistique, datée de 1985, portant sur la Bibliothèque nationale de Turin. Elle est pourtant éloquente : il en résultait, en effet, qu'un manuscrit n'avait été consulté, en moyenne, qu'une fois tous les quatre ans et que la durée de consultation moyenne n'était que d'un jour. 37% des volumes n'avaient jamais été consultés pendant une période de cinq ans, et 37% n'avaient été consultés qu'une fois. Donc, les ¾ du patrimoine avaient été extraits des magasins au plus une seule fois en cinq ans. Seuls 1% des manuscrits avaient été consultés plus d'une fois par an.⁸

Cela dit, nul ne conteste l'idée selon laquelle la numérisation des manuscrits pourrait réduire la nécessité d'accéder directement à l'objet original. En effet, il y a deux situations qui donnent lieu à des consultations, alors qu'elles pourraient être évitées. La première est celle où l'on connaît l'existence d'un manuscrit potentiellement « intéressant » — ou pertinent a priori dans le cadre d'une recherche quantitative — mais où l'on ne sait rien ou presque sur lui. Dans ce cas, il arrive qu'on consulte à l'aveuglette, simplement pour

⁵ La bibliothèque en question était... la Bibliothèque Laurentienne. On doit néanmoins souligner que cet épisode ancien ne préjuge en rien du présent et de l'avenir : la politique concernant l'accès aux originaux présente, en effet, beaucoup de variations, dans la diachronie et dans la synchronie, en fonction du point de vue du personnel dirigeant, ce qui introduit une variable aléatoire extrêmement dommageable. Un autre passage de ce petit article pourrait avoir été écrit aujourd'hui : « On a souvent l'impression qu'aujourd'hui, dans l'utilisation des biens culturels, prévaut une politique qui tend à s'en servir comme d'un œillet à la boutonnière ou d'un moyen de publicité, plutôt que comme une source éternelle d'enrichissement de notre humanité et comme le témoignage irremplaçable d'un passé que nous devons étudier pour qu'il revive en nous dans toute sa richesse. Il n'est pas rare que les choix culturels d'une administration locale, d'une bibliothèque, d'un musée, et même d'un ministère n'aient pas, en dernière analyse, des motivations bien différentes de celles d'une grosse industrie ou d'une banque qui patronnent une restauration ou une publication » (Rizzo 16–17).

⁶ Ce terme doit être entendu au sens large : dans un objet archéologique, on peut « lire » autre chose que le texte.

⁷ L'article 4 mérite également d'être cité : « Toute restriction apportée à la consultation du manuscrit ne peut être fondée que sur la nécessité impérieuse d'en assurer l'intégrité *en vue des études à venir* [souligné par moi]. Sauf pour un temps limité, le manuscrit ne sera jamais privé totalement de sa finalité première, qui est d'être lu » (2).

⁸ Voir Vitale Brovarone, qui remarque d'ailleurs, à juste titre, que la consultation d'un manuscrit peut permettre de mettre en évidence les symptômes de certains processus de dégradation qui passeraient inaperçus dans le cas contraire.

vérifier à quel type d'objet, de texte ou d'image l'on a affaire. La seconde est celle où l'on doit transcrire ou collationner entièrement le texte d'un ouvrage, ou encore analyser un programme iconographique dans son ensemble. Dans le premier cas, la consultation *de visu* pourrait être évitée ; dans le second, elle demeure de toute manière nécessaire, mais sa fréquence pourrait baisser radicalement du fait qu'une bonne image en couleurs fournit tout de même une quantité non négligeable de détails. Il s'agirait, quoi qu'il en soit, d'un effet automatique, et non d'une contrainte imposée en échange de l'accès aux images : personne n'est assez stupide pour échanger définitivement un partenaire en chair et en os avec une poupée gonflable, aussi ressemblante soit-elle.

Ces réserves étant formulées, il serait bon que les chercheurs établissent une sorte de « charte de la numérisation » définissant les critères minimaux pour que les résultats de l'opération correspondent réellement à leurs aspirations. Pour ma part, j'insisterais surtout sur les points suivants :

1. La numérisation doit se faire directement sur les originaux, et non à partir de microformes.
2. Pour chaque bibliothèque, l'opération doit concerner à terme la totalité des fonds, et la totalité des manuscrits de chaque fonds.
3. Chaque livre doit être intégralement numérisé.
4. Toutes les pages doivent être numérisées avec la même définition.⁹
5. La totalité des images doit être librement et gratuitement disponible sur le Web.
6. Les images doivent être disponibles sur le Web sous une forme exploitable. Cela signifie qu'elles doivent pouvoir être agrandies sans « pixelliser », afin de pouvoir mettre en évidence tous les détails de la page écrite (structure du parchemin, piqûres, réglure, ductus de l'écriture, etc.) et ce, quelles que soient les dimensions du volume original. En d'autres termes, puisque cela est techniquement possible, il faudrait, à certains égards, que l'on puisse examiner une page dans des conditions plus favorables que celles dont on jouit lors de la consultation directe en salle de lecture.
7. Les images doivent pouvoir être aisément comparées et retravaillées à l'intérieur d'un logiciel de retouche. Cela présuppose qu'elles puissent être téléchargées à loisir, soit isolément, soit, éventuellement, en bloc (par exemple comme fichier PDF), et qu'elles aient été créées dans un format universellement reconnu.
8. Les images doivent pouvoir être librement utilisées et reproduites dans le cadre de la normale activité scientifique, pourvu que les règles concernant la propriété intellectuelle (notamment l'obligation de citation) soient rigoureusement respectées.

Combien, parmi les entreprises achevées, en cours ou en projet, satisfont toutes les exigences exprimées ci-dessus ? Il n'y a pas lieu, dans le cadre de cette contribution,

⁹ Ainsi, les manuscrits de la Bibliothèque du *Sacro convento* d'Assise ont été numérisés selon deux standards différents : « I parametri qualitativi indicati dalla Biblioteca Digitale Italiana, la risoluzione digitale adottata va dai 300 ai 600 dpi per i manoscritti non miniati (da 400 a 1000 dpi per i miniati) ».

de distribuer de bons ou de mauvais points, mais il serait tout de même intéressant de passer au crible de ces critères tout ce que l'on trouve à l'heure actuelle sur le Web.¹⁰ Je ne crois pas me tromper en affirmant que très peu d'initiatives résisteraient à ce test.

3. Numérisation, catalographie et bases de données

Toutefois, à supposer même que tous les fonds soient numérisés avec des standards satisfaisants et que toutes les images soient diffusées auprès de la communauté scientifique dans les meilleures conditions, il n'en reste pas moins que les albums d'images ne sauraient être livrés sans le moindre support descriptif ; et c'est sans doute ici que le bât blesse.

Ce n'est pas pour rien, soyons-en sûr, que la numérisation des fonds a été d'emblée déconnectée de toute initiative dans le domaine de la catalographie. Compte tenu de l'état d'avancement du catalogage du patrimoine manuscrit dans certains pays et de son état prévisible dans l'avenir proche et lointain, c'était le seul moyen de faire démarrer rapidement la nouvelle démarche et de l'achever dans des délais raisonnables.

Face à l'absence d'un support descriptif conçu *ad hoc*, la solution généralement adoptée consiste à numériser les catalogues imprimés déjà parus et à instituer un lien hypertextuel entre les images et la notice correspondante. Or, ces catalogues sont souvent anciens, et parfois très anciens : dans le cas de la Bibliothèque Laurentienne, ils remontent au XVIII^e siècle. De plus, bien des bibliothèques — et non des moindres — ne possèdent même pas de catalogue imprimé de tout ou partie de leurs fonds.

Or, il est évident que cette situation ne saurait perdurer et ce, pour deux raisons.

À partir du moment où le nombre de livres numérisés atteint un niveau conséquent, toute navigation efficace dans le corpus devient impossible si l'on ne dispose pas *au préalable* de quelques données de base sur chaque volume. À titre d'exemple, supposons que, dans la base numérisée d'une grande bibliothèque, on veuille sélectionner l'ensemble des bibles dites « de poche ». Puisque les manuscrits de la Bible sont nombreux et leurs typologie très variée, une simple liste très laconique de « Biblia sacra » obligera à cliquer sur chaque élément pour voir les images et établir de quoi il s'agit. Dans ce cas précis, il est évident que la présence des dimensions dans le résultat de la requête initiale serait la bienvenue, ce qui permettrait de ne pas tenir compte des manuscrits dont les dimensions sont supérieures à une certaine limite. Dans d'autres cas, on pourrait avoir besoin de la mention du support ou du nombre de feuillets, ou de la langue du texte. On voit bien qu'à terme on ne saurait faire l'économie d'une base de données, même très squelettique, où seraient enregistrées ces quelques données fondamentales.

¹⁰ On pourra en trouver la liste sur le portail *Digitale Handschriften*. Pour une description raisonnée de chaque initiative, voir les commentaires de Marc Smith dans le portail *Menestrel*.

Par ailleurs, un simple lien entre les images et le texte d'un catalogue, même si ce dernier était récent et les notices extrêmement fouillées, ne constituerait que le prolongement de la situation imposée par la rigidité de la forme imprimée, avec l'avantage toutefois de pouvoir disposer d'images en couleurs. Or, cette rigidité empêche d'établir une liaison automatique entre deux ou plusieurs notices et entre deux ou plusieurs jeux d'images relatives à des objets différents.

Il en serait de même si, à partir des images numérisées, on construisait *ex nihilo* un catalogue calqué sur le modèle des catalogues imprimés. On conviendra aisément, en effet, que, dans ce domaine, l'apport du Web ne saurait se réduire à la production à bas coût d'un catalogue, certes illustré, mais irrémédiablement figé. En d'autres termes, pour donner une nouvelle vie à ce qu'on appelle couramment « catalogue », il faudrait que celui-ci soit structuré comme une base de données.

Avant d'aborder ce problème, il convient de se pencher sur l'interrelation étroite qui lie l'existence d'images numérisées et la démarche descriptive.

Lorsque j'avais envisagé l'existence d'une *Bibliotheca universalis*, l'une des objections à ce projet « chimérique » mettait en exergue une opposition de fait entre numérisation et catalogage. La véritable procédure cognitive étant représentée par le catalogage, il était à craindre que l'appât des images et le différentiel de rapidité entre les deux entreprises conduise inévitablement à détourner les financements vers la numérisation, sacrifiant ainsi le savoir au profit du « voir ». En d'autres termes, les images numérisées auraient fini par constituer une sorte de palliatif à l'absence de catalogue ; absence qu'elles auraient de fait contribué à pérenniser.

Ce point de vue n'est pas tout à fait injustifié, mais il a le défaut d'être statique. Il faut réfléchir, en effet, sur le fait qu'un catalogue est toujours porteur de deux fonctionnalités différentes : d'un côté, il est un *médiateur descriptif* entre l'objet et ceux qui, en étant éloignés, se trouvent dans l'impossibilité d'y accéder ; de l'autre, il est un *médiateur cognitif*, en ce sens qu'il fournit des informations complexes dont l'acquisition ne serait pas immédiate même en présence de l'objet, du fait qu'elle requiert des compétences spécifiques et souvent pointues.

La disponibilité des images sur le Web a comme effet de changer la donne, en ce sens qu'une bonne partie de la médiation descriptive n'est plus strictement nécessaire à partir du moment où les pages du manuscrit sont devant nos yeux. Dès lors, la tâche ingrate qu'est le catalogage s'en trouve considérablement allégée. Bien entendu, cette affirmation optimiste doit être nuancée : elle n'est valable que pour les utilisateurs « séquentiels » ; ceux qui, comme c'est le cas le plus fréquent, se contenteraient de prendre connaissance de chaque manuscrit pris singulièrement et de naviguer dans la *Bibliotheca* à partir de quelques paramètres essentiels (date, localisation, support, etc.). Cela dit, le conflit de priorité entre la numérisation et le catalogage peut être atténué en étalant le catalogage dans le temps et en le distribuant dans l'espace. Le dispositif

qui permettrait de répondre simultanément à ces critères est celui que j'appellerais volontiers « catalogue grand-ouvert ».

Nous connaissons déjà le concept de « catalogue ouvert » dont l'exemple autoproclamé le plus connu est le *Catalogo aperto* de la Bibliothèque Malatestienne de Cesena.¹¹ Ce catalogue ouvert consiste en un ensemble fait de notices descriptives, des pages entièrement numérisées de chaque manuscrit décrit, ainsi que de contributions scientifiques — anciennes, récentes ou nouvelles — relatives au fonds manuscrit de la bibliothèque, numérisées sous des formes variées. L'ensemble ainsi défini est dynamique, puisqu'il est sans cesse augmenté, et les notices descriptives ne sont pas figées dans le marbre.

Un catalogue « grand-ouvert » est, lui, quelque chose de plus : les notices sont structurées en base de données et comportent donc des champs de recherche plus ou moins détaillés ; mais, surtout, l'ensemble est géré de manière interactive. En effet, grâce à la présence des images, le rôle du catalogue en tant que médiateur descriptif comporte désormais deux aspects qui, sans être différenciés sur le plan épistémologique, peuvent subir, dans la pratique, deux traitements différents : il y a d'un côté les caractéristiques qui sont visibles sans ambiguïté sur l'image numérisée et, de l'autre, celles qui ne le sont pas. On aboutirait ainsi à une tripartition ergonomique qui serait transversale par rapport à la stratification conceptuelle qui, dans une description traditionnelle, distingue les aspects matériels, les aspects textuels et les aspects historiques.

Dans cette nouvelle perspective, toutes les informations, quelle que soit leur nature, que l'on peut acquérir par la simple observation visuelle n'ont plus besoin d'être explicitées (l'utilisateur les acquiert lui-même) ou peuvent être insérées directement dans le catalogue/base de données sur le Web sans nécessité de qualification scientifique particulière. Celles qui sont tout aussi évidentes à relever, mais qui requièrent un examen de l'objet en chair et en os, peuvent l'être *in loco* par les bibliothécaires, les chercheurs ou un personnel recruté temporairement à cet effet. Enfin, tout ce qui relève d'un travail d'expertise — datation, localisation, identification de copistes, d'artistes ou de possesseurs — ne peut être réellement pris en charge que par des chercheurs expérimentés dont la compétence porte sur des périodes, des aires géographiques et des typologies textuelles spécifiques et ce, quel que soit le milieu où ils travaillent (bibliothèques, universités, institutions de recherche). Bien entendu, l'existence d'un catalogue « grand-ouvert » en cours d'avancement ne doit pas empêcher d'utiliser, faute de mieux et provisoirement, les données provenant des catalogues existants.

¹¹ Cartelli *et al.* Un autre exemple de « catalogue ouvert », différent dans son fonctionnement de celui de la Bibliothèque Malatestienne, est constitué par la *Nuova Biblioteca Manoscritta*. Cette initiative est consacrée aux bibliothèques de la Vénétie et rassemble surtout du matériel d'époque moderne (Bernardi *et al.*). Il s'agit en fait d'un catalogue collectif où la saisie et la mise à jour interactive des données sont effectuées par les divers collaborateurs directement sur le Web.

On comprendra donc que des paramètres immédiatement « catégorisables » ou quantifiables, tels que le support, la disposition du texte, le nombre de lignes par page, la présence de réclames ou de titres courants, les rubriques, les incipits et ainsi de suite, peuvent ne pas être jugés prioritaires (mais dans ce cas ils ne pourront être la cible d'un formulaire de « recherche avancée ») du fait que, quel que soit leur intérêt dans le cadre d'enquêtes particulières, ils ne fournissent pas, comme c'est le cas pour la datation et la localisation, d'informations indispensables à la grande majorité des utilisateurs. On peut tout aussi bien concevoir, inversement, que les informations indispensables, mais qui malheureusement sont bien souvent fondées sur une évaluation synthétique subjective, fassent l'objet d'appréciations divergentes qui pourraient être rendues publiques dans un forum de discussion couplé à chaque manuscrit, afin que les utilisateurs de la base puissent en tirer profit et, le cas échéant, les mettre en balance.

4. Les bases de données en ligne : peuvent mieux faire. . .

Il n'y a pas lieu d'entrer dans les détails de la structure et du fonctionnement de ce type de catalogue dans le cadre de la présente contribution. Ce qui importe de souligner, c'est en revanche la contradiction dont tout catalogue de manuscrits disponible sur le Web est inévitablement porteur.

Malgré les difficultés inhérentes à leur création et à leur exploitation, les bases de données conçues à l'époque de l'informatique « lourde » possédaient un certain nombre de qualités. Puisqu'elles étaient destinées au calcul statistique, elles avaient le plus souvent une structure tabulaire dont les lignes — correspondant à des « individus » — constituaient les enregistrements et les colonnes les champs. L'information relative à chaque individu était rigoureusement catégorisée en « objets » ; chaque objet était pourvu de plusieurs propriétés distinctes, et chaque propriété présentait plusieurs modalités rigidement codifiées. Par ailleurs, chaque modalité de chaque propriété était rigoureusement codifiée de la même manière.

C'étaient-là des évidences pour tout concepteur de bases de données. Malheureusement, de ce point de vue, la situation a beaucoup empiré depuis. En effet, les bases de données disponibles sur le Web n'obéissent que très rarement à ces principes. Elles ne sont pas destinées au calcul statistique (et ne sont donc pas confrontées à ses verdicts impitoyables), mais à fournir une réponse adéquate à des interrogations — des requêtes — provenant d'utilisateurs tiers ; et une réponse n'est jamais le résultat organisé d'un calcul, à savoir une statistique descriptive : il s'agit toujours d'une sélection, c'est-à-dire d'une liste des individus de la base dont les caractéristiques correspondent aux paramètres de la requête. La sélection est l'œuvre d'un « moteur de recherche » qui utilise des opérateurs booléens d'une manière plus ou moins complexe, selon que la recherche peut ou non englober simultanément plusieurs champs et que les opérateurs

booléens élémentaires peuvent faire l'objet de combinaisons plus ou moins élaborées. Une fois la réponse obtenue, un simple clic permet de visualiser (copier, imprimer) l'ensemble de l'information qui concerne un ou, dans le meilleur des cas, plusieurs individus préalablement retenus dans la sélection.

Dans cette nouvelle perspective, ce que l'on peut uniquement prétendre d'une base, c'est la pertinence de la sélection obtenue par rapport à la requête envoyée. Cela revient à dire que la sélection doit contenir *tous* les individus concernés par la requête et *aucun* intrus.

Cet objectif est d'autant plus facile à atteindre que la structure de la base remplit au mieux les critères définis ci-dessus. Or, plus une structuration est performante du point de vue analytique, moins elle est satisfaisante du point de vue synthétique. Autrement dit, ce que l'on gagne en précision dans la requête est perdu en visibilité dans la réponse. Pour s'en rendre compte, il suffit de parcourir la description du système de codification TEI (*Text Encoding Initiative*) relatif à la description des manuscrits :¹² on voit nettement que, à partir d'une description sommaire provenant d'un catalogue imprimé, la structuration de l'information peut se faire à des niveaux de plus en plus détaillés dont le plus simple est celui qui permet de reproduire tel quel le texte d'origine et le plus détaillé correspond à un état de l'information extrêmement « décomposé ». Bien évidemment, si l'on souhaite maximiser les performances du moteur de recherche, le balisage du texte en champs doit non seulement être très fin, mais également s'accompagner d'une normalisation stricte du contenu de chaque champ, ce qui implique la perte du tissu discursif, dont la réception est aisée et immédiate, au profit d'une présentation sèchement énonciative.

Le travail de structuration, lorsqu'on le pousse à l'extrême, est assurément très onéreux, surtout lorsque la base n'est pas créée *ex nihilo* mais reconfigure l'information puisée dans un matériel préexistant. De plus, le résultat va à l'encontre des attentes de l'utilisateur moyen : celui-ci, en effet, aime retrouver sur l'écran ce qu'il a d'ordinaire devant les yeux lorsqu'il parcourt les pages d'un catalogue imprimé ; il s'accommode donc mal d'une présentation excessivement structurée, jugée trop « décharnée » et perçue, par conséquent, comme dépayssante.

Les solutions adoptées par les concepteurs de bases de données sur le Web représentent en général un compromis entre les deux exigences opposées. Toutefois, il arrive souvent que la structuration en champs se résume à la simple stratification typologique de l'information, en l'absence de toute formalisation — et *a fortiori* de toute normalisation — à l'intérieur de chaque champ. La notice ci-contre, tirée du catalogue des manuscrits enluminés de la British Library, disponible sur le Web, constitue un bon exemple de cette manière de procéder : son contenu ne diffère en rien de ce qu'on pourrait lire

¹² Ce système de codage n'est cité ici qu'à titre d'exemple ; il ne s'agit en aucun cas d'en évaluer les qualités intrinsèques.

sur une page d'un bon catalogue imprimé, et la mise en paragraphes est à peine plus accentuée.

Cependant, comme on peut le voir dans la capture d'écran, ce catalogue est également une base de données que l'on peut interroger grâce à un formulaire de « recherche avancée » portant sur un certain nombre de champs dont quelques-uns sont indexés. Lorsque les informations dans la base ne sont pas rigoureusement structurées, le moteur de recherche opère une exploration contextuelle du champ concerné par la requête (recherche « full text »). C'est une sorte de « pêche à la ligne » dont le résultat dépend à la fois du degré d'élaboration du moteur, du degré de promiscuité des « objets » et de leurs propriétés à l'intérieur du même champ, et, *last but not least*, de la variabilité des formulations employées dans la base pour définir une même modalité. En général, le mélange de ces trois sources potentielles d'inconvénients engendre un résultat décevant : la sélection obtenue pêche par défaut et/ou par excès, et de plus certaines caractéristiques dont la formulation dans la base n'est pas assez rigoureuse ne peuvent faire l'objet d'aucune requête ayant une chance quelconque de succès.

Le catalogue/base de données en question — qui au demeurant peut rendre d'excellents services — n'échappe pas à ce genre de travers, et malheureusement il n'est pas le seul.

Le tableau ci-dessous (tableau 2) montre que le résultat de certaines requêtes diverge selon que la requête est effectuée dans le formulaire de recherche avancée ou à partir des index correspondants. La recherche avancée semble ignorer toutes les subdivisions de l'Italie, sauf l'Italie centrale. « Italy, Central, Rome » et « Rome » fournissent quatre résultats différents. L'index comporte deux entrées différentes pour le nord-est de l'Italie. L'expression « Italy or France » donne lieu à une sélection de 1887 manuscrits, alors que « Italy » + « France » prises séparément donnent lieu à 1972 manuscrits (1895 par l'index). Enfin, dans la requête « Italy or France », « or » est systématiquement interprété comme un opérateur booléen ; il est donc impossible de sélectionner les volumes désignés dans la base par l'expression « Italy or France » lorsqu'elle signifie qu'il y a doute sur le lieu d'origine.¹³

Dans certains cas, l'obstacle pourrait sans doute être contourné grâce à un usage plus sophistiqué des opérateurs booléens. Mais bien souvent le moteur de recherche, trop fruste, interdit toute subtilité logique. Par ailleurs, l'utilisateur ordinaire est le plus souvent incapable de mettre en œuvre des manipulations savantes, d'autant que, s'il ne procède pas à des requêtes comparatives, il ne peut même pas s'apercevoir de l'existence d'un problème.

¹³ L'effectif de la base étant en augmentation constante, il convient de préciser que la situation décrite dans le tableau correspond à des requêtes effectuées le 15 mai 2010.

Author	Pseudo-Augustine ; John of Fécamp
Title	Meditationes, Soliloquia (ff. [28–74v]), and Manuale (ff. [74v–97v]); Liber Supputationum (suspiria) (ff. 97v–[114]); and various prayers or meditations
Origin	Netherlands, N.
Date	2 nd half of the 15 th century
Language	Latin
Script	Gothic
Decoration	5 large initials in blue with penwork decoration and colour washes with leaves and acorns, extending to form a partial border at the beginning of texts and major divisions (ff. [5], [28], [75], [98], [114]). Large initials in red. Paraphs in red. Highlighting of letters in red. Rubrics in red. Spaces left for initials.
Dimensions in mm	115 × 75 (85 × 60)
Official foliation	Unfoliated (ff. [148])
Form	Parchment codex
Binding	Pre-1600. 15 th -century Netherlandish brown calf over wood boards, with blind tooling including of the letter 'T', a dragon and leaves, and remnants of 1 pair of clasps
Provenance	The Cistercian abbey of Mount St Bernard, near Leicester : its book-plate, 'Ex Libris Abbatiae De Monte Sti. Bernardi.' (inside upper cover). Mrs. Constance Goetze, wife of the popular painter Sigismund Goertze (d. 1939) and former owner of a number of manuscripts now at the Fitzwilliam Museum ; sold at Sotheby's, London, 2 December 1942, lot 326. Henry Davis (b. 1897, d. 1977), businessman and book collector : his book-plate (inside upper cover) ; and manuscript number-plate 'HD M27' (first folio). The Henry Davis Gift of book-bindings was donated to the British Museum in 1968.

TABLE 1. Manuscrits enluminés de la British Library. Description du ms Henry Davis Collection 597.

CATALOGUE OF ILLUMINATED MANUSCRIPTS

About | Simple search | Manuscript search | Advanced search | Tours | Glossary | Contact us | Main

Advanced search

print home site search back

search tips new search modify search

Author: Illumination:
 Contents: Language:
 Place of Origin: Provenance:
 Dated between: and Script:
 Dated / datable: ☐ Format:
 Composite codex: ☐ Binding:
 Scribe: Collection:
 Artist: Bibliography:

Image description: (Note that some images on this site do not have captions or descriptions)

Browse Indexes of

Places of origin
 Scripts
 Scribes
 Artists

Champs indexés

This page contains names, dates, language, and other terms from the detailed records. You can search on one type of information or combinations. For full information on how to search see [Search tips](#). For information on the different fields see [About the records](#).

FIGURE 1. Manuscrits enluminés de la British Library : écran « recherche avancée ».

Cela dit, il vaut mieux réserver à une meilleure occasion l'illustration approfondie des inconvénients engendrés par les champs « fourre-tout » et la mise en garde quant aux aléas imprévisibles qui accompagnent inévitablement l'exercice difficile de la recherche contextuelle ; aléas qui, d'ailleurs, ne relèvent nullement du... hasard mais sont, bien au contraire, l'expression d'une conception plutôt « laxiste » des bases de données dont les causes sont multiples, et les effets sous-estimés.

L'idéal, ce serait que l'on puisse concilier le besoin d'une présentation « traditionnelle », plus conforme à ce qu'attendent en général les chercheurs de nos disciplines, avec la possibilité de mettre en œuvre des requêtes plus performantes et dont les résultats seraient plus fiables. Cela ne pourrait se faire que si l'information enregistrée sous une forme familière à la majorité des utilisateurs, et donc destinée à l'écran, était couplée avec une base de données sous-jacente dotée d'une structure plus rigide mais plus robuste qui, elle, pourrait fournir des réponses à un moteur de recherche élaboré. Il ne semble pas, hélas, que l'on s'oriente dans cette direction ; et non seulement parce que cette tâche monopoliserait beaucoup d'énergie, mais aussi parce que la formalisation des phénomènes les plus complexes, tels que la décoration, présente des difficultés

requête	Recherche par l'index	Recherche avancée
Italy	863	911
Italy, N.	222	911
Italy N. E.	12	911
Italy, N. E.	131	911
Italy, N. W.	7	911
Italy, S.	30	911
Italy, Central	293	296
Italy, Central, Rome	3	69
Rome	80	84
France	1032	1061
Italy or France	nd	1887
Italy and France	nd	21

TABLE 2. Manuscrits enluminés de la British Library : quelques résultats de requêtes¹⁴.

conceptuelles difficiles à surmonter qui peuvent paraître rédhibitoires dès lors qu'on n'entrevoit pas clairement la nécessité de les affronter.

5. L'interconnexion des catalogues et des bases de données

Cependant, même si la numérisation du patrimoine manuscrit se faisait selon les vœux des chercheurs ; même si elle était associée à des bases de données solides et fiables, issues de l'élaboration progressive d'un catalogue « grand-ouvert » collectif et interactif, et aptes, par conséquent, à faciliter la navigation dans les « magasins » virtuels d'une bibliothèque, nous serions encore loin d'avoir affaire à une véritable *Bibliotheca universalis*.

Ce que nous aurions, ce serait une liaison bidirectionnelle hautement perfectionnée, une sorte de télévision où la souris, telle une télécommande, nous permettrait de « zapper » très rapidement et d'une manière efficace d'un manuscrit à l'autre à l'intérieur d'une seule et même bibliothèque. Or, la télévision a été inventée il y a bien longtemps et le Web représente quelque chose de profondément différent du fait de sa structure en réseau, car il permet, en théorie, d'accéder simultanément à des informations puisées dans des sources multiples et de les rassembler sous une forme exploitable.

Pour illustrer la portée de ce propos, songeons à deux opérations qui seraient largement profitables à la démarche érudite : la recherche des *membra disiecta* —

¹⁴ Sélection obtenue en cliquant sur le champ indexé. « Recherche avancée » = sélection obtenue en insérant la même chaîne de caractères dans le champ approprié du formulaire « recherche avancée ».

à savoir la reconstitution de manuscrits autrefois dépecés et dispersés aujourd'hui aux quatre vents — et la reconstitution des bibliothèques anciennes à partir des premiers et derniers mots du deuxième et de l'avant-dernier feuillet des manuscrits inventoriés. Il est évident que ce type de recherche serait accéléré d'une manière décisive s'il existait un moteur capable de puiser ces informations dans les bases de données de toutes les bibliothèques et de les fournir rassemblées à l'utilisateur.

Comment mettre en œuvre des dispositifs permettant d'obtenir ce résultat ? Bien sûr, il n'est guère besoin que l'information contenue dans toutes les bases soit strictement uniformisée. L'uniformisation ne devrait concerner que les données « élémentaires » et/ou directement quantifiables. Pour le reste, il suffirait de créer une ou plusieurs « métabases intégratives ». Une métabase intégrative n'est rien d'autre qu'une liste d'autorités pourvues de toutes leurs variantes attestées. Le choix de la forme faisant autorité n'a aucune importance : à la limite, on pourrait se contenter d'un simple numéro. Un tel dispositif est techniquement très facile à réaliser et pourrait être appliqué aux champs qui, dans une hypothétique *Bibliotheca universalis*, constitueraient à coup sûr les cibles de la très grande majorité des requêtes : l'auteur et le titre.

Sur le plan scientifique, la situation apparaît plus compliquée. Si de telles listes d'autorités sont proposées depuis longtemps aux bibliothécaires de plusieurs pays afin de faciliter le catalogage, il n'existe, à ma connaissance, qu'une seule base publique potentiellement intégrative : le *Thesaurus* du *Consortium of European Research Libraries* (CERL) qui, cependant, concerne essentiellement le livre imprimé.¹⁵ Lorsqu'on introduit dans le formulaire de recherche le nom d'un auteur — par exemple « Thomas de Hibernia », l'auteur du *Manipulus florum* — la forme vedette s'affiche, suivie de toutes les variantes repérées :

Headings : Thomas <Palmeranus>

Variant names :

Hibernicus, Thomas

Hibernia, Thomas de

Hybernicus, Thomas

Palmer, Thomas

Palmeranus, Thomas

Thomas <Hibernicus>

Thomas <Hybernicus>

Thomas <Palmer>

Thomas <Palmerstonensis>

Thomas <Palmerstonus>

Thomas <aus Palmerston> [(VD-16)]

¹⁵ Pour les arts plastiques, il faut signaler la base *Union List of Artist Names Online* (ULAN) du Getty Museum.

Thomas <d'Irlande>
 Thomas <de Hibernia>
 Thomas <de Hybernia>
 Thomas <de Palmerstown>
 Thomas <of Ireland>
 Thomas <von Irland>

Pour pouvoir sélectionner tous les manuscrits des œuvres composées par l'auteur du *Manipulus florum*, il suffirait que chacune des variantes repérées dans les sources primaires (manuscrits) et secondaires (catalogues) pointe systématiquement vers « Thomas <Palmeranus> » dans la métabase.

Trop simple ? Sans doute, car, si la requête était envoyée à partir de « Thomas Hibernicus » plutôt qu'à partir de « Thomas de Hibernia », la sélection des auteurs-cibles ne se limiterait pas à « Thomas Palmeranus » ; en effet, dans le *Thesaurus* du CERL, ce nom renvoie également à un franciscain mort en 1270, ainsi qu'à un « Thomas <aus Palmerston> » qui, ayant vécu au XVI^e siècle, n'a rien à voir avec l'auteur du *Manipulus florum*. Le même problème se poserait si la requête portait sur « Thomas <Palmer> », car ce nom est aussi celui d'un dominicain qui a vécu entre 1371 et 1413.¹⁶

Les cas d'homonymie étant assez nombreux dans la littérature médiévale, on comprend qu'il est impossible d'échapper à un travail préalable de désambiguation, faute de quoi la recherche se révélerait particulièrement hasardeuse : si une même variante peut renvoyer à deux ou plusieurs autorités différentes, il est nécessaire que les correspondances correctes soient établies au moment de la création de la base.¹⁷

Quelle est la situation actuelle sur le terrain ? Il existe une base informatisée — *Manuscripta Mediaevalia* — qui regroupe 63000 manuscrits issus de catalogues imprimés concernant tout particulièrement les régions germanophones. Une requête sur l'auteur du *Manipulus florum*¹⁸ a abouti aux résultats suivants :

Aucune sélection n'englobe la totalité des manuscrits du *Manipulus florum* contenus dans la base — qui sont vraisemblablement au nombre de 37 —, ce qui signifie qu'aucune sélection n'est un sous-ensemble de la sélection la plus riche. Ce tableau sommaire n'a pour but que de mettre l'accent sur les difficultés méthodologiques et pratiques de toute entreprise d'intégration qui rassemble des sources préexistantes d'origine disparate ; le terme « disparate » devant s'entendre comme « élaboré par des personnes différentes »,

¹⁶ La base du CERL renvoie également à d'autres homonymes qui toutefois, ayant vécu après le XVI^e siècle, peuvent être négligés a priori. Sous la vedette « Thomas Palmeranus », elle met d'ailleurs en garde contre les confusions possibles avec ses homonymes connus : « Nicht identisch mit Thomas <Hibernicus> (OFM ; gest. 1275) und Thomas <Palmer> (OP ; um 1319/ 1413), mit denen er bisweilen gleichgesetzt wurde ».

¹⁷ Ce travail est encore plus nécessaire — et surtout beaucoup plus lourd — lorsqu'il s'agit de noms de lieu, car les cas d'homonymie y sont extrêmement nombreux.

¹⁸ À savoir : Thomas de Hibernia (ou autre) UND *Manipulus florum*, car Thomas d'Irlande a écrit autre chose que le *Manipulus*, et d'autres auteurs ont écrit des *Manipulus florum*.

type de recherche	requête	réponses
recherche standard	Thomas de Hibernia	31
	Thomas Hibernicus	26
	Thomas Palmeranus	30
	Thomas Palmer	24
	Thomas Hybernicus	24
	Thomas de Hybernia	24
recherche avancée	Thomas <Palmeranus>	21
	Palmeranus	23
	Thomas de Hibernia	5
	Hibernicus	25

TABLE 3. Manuscripta mediaevalia : résultats de la recherche sur Thomas d'Irlande.

même dans le cadre d'une entreprise organisée. Les responsables de la base *Manuscripta mediaevalia* se montrent d'ailleurs tout à fait conscients de ces difficultés.¹⁹

6. Les obstacles extrascientifiques

Le parcours conduisant à la réalisation d'un arrière-plan scientifique solide, apte à maximiser les performances d'une *Bibliotheca universalis*, est donc parsemé d'embûches, ce qui n'est pas étonnant et, à la limite, pourrait même paraître stimulant. Quoi qu'il en soit, on peut prévoir, sans risque de démenti, que ce parcours ne peut être que long : il s'agit, en effet, d'un travail critique patient et obscur qui doit être assumé par des personnes hautement qualifiées et ce, dans un domaine dont l'utilité sociale n'est évidente qu'aux yeux de quelques passionnés. Or, ce qui est à la fois cher, obscur et « inutile » a toutes chances d'être relégué au plus bas dans l'échelle des

¹⁹ « Bitte beachten Sie bei Ihren Recherchen, dass die Qualität der Daten und die Systematik der Register in den einzelnen Katalogen sehr unterschiedlich sein kann. Diese Varianzen spiegeln sich trotz aller Bemühung um Normalisierung auch in der Datenbank. Die Zentralredaktion arbeitet kontinuierlich an der Klassifizierung der Registereinträge nach Personen- und Körperschaftsnamen, Orten, Datierungen, Werktiteln, Sachschlagworten etc., allerdings ist die Datenmenge schneller gewachsen als diese Arbeiten voranschreiten konnten. Daher sind (noch) nicht alle Daten über die Suche in den inhaltlich spezifizierten Indexfeldern zu finden, wohl aber – mit etwas Phantasie – in der Freitextsuche im "Gesamtregister" ». Notons par ailleurs que certains programmes de catalogage et/ou de numérisation ont créé en interne leur propre métabase de noms d'auteurs : c'est le cas pour *e-codices*, site consacré à la numérisation des manuscrits des bibliothèques suisses. Le fait que ce programme concerne jusqu'à présent un nombre limité de manuscrits et que le travail de normalisation peut donc être effectué au fur et à mesure de l'avancement de l'entreprise facilite bien évidemment les choses.

priorités, dans une dynamique négative qui présente depuis quelque temps des signes indiscutables d'accélération, du moins sur le plan institutionnel (suppression de chaires, marginalisation des « sciences auxiliaires » dans les plans d'études des universités).

Comme si cela ne suffisait pas, il faut également tenir compte d'obstacles d'une toute autre nature qui, je le crains, vont se révéler encore plus redoutables. Celui qui s'impose avec le plus d'évidence a trait à l'emprise presque « obsessionnelle » des copyrights éditoriaux sur la diffusion de la littérature scientifique.

Dans la mesure où la quasi totalité des catalogues de manuscrits se présente actuellement sous la forme imprimée, leur visualisation et leur utilisation sur le Web en tant que support normalisé pour les notices descriptives des manuscrits appartenant à des bases interconnectées suscitera inévitablement l'hostilité des éditeurs dont les droits — rappelons-le — couvrent une période de 70 ans. Si l'on songe au fait que bon nombre de ces publications à faible tirage n'ont pu voir le jour sans le recours à des subventions publiques, on peut s'étonner de ce que, comme l'impose la loi, les droits demeurent en vigueur même si les coûts de production ont été largement amortis et l'éditeur n'envisage aucune réimpression dans un avenir prévisible. Il s'agit, en fait, d'un véritable « droit de glaciation » qui de plus, dans la plupart des cas, ne profite à personne.

Bien entendu, les ayant droit s'opposeront également avec succès à toute initiative visant à diffuser gratuitement sur le Web la littérature scientifique. En réalité, il serait temps de comprendre que, dans le domaine de la recherche, en dehors de quelques opérations d'envergure où l'intervention de professionnels hautement qualifiés se révèle indispensable, la figure de l'éditeur commercial n'a plus de raison d'être. Autrefois indispensable, encore utile dans un passé récent, le système de production éditorial est désormais devenu nuisible, dans la mesure où il n'est plus un moteur, mais un obstacle au processus de diffusion des connaissances.²⁰ On ne peut rester indifférent au fait que, si une contribution scientifique était librement consultable sur le Web, le nombre de ses lecteurs serait multiplié par cent, voire par mille ; et la finalité d'une contribution scientifique n'est pas de rémunérer un investissement lourd, mais bien de donner la plus grande diffusion à la connaissance dans les plus brefs délais avec une dépense minimale d'argent et d'énergie.

Aujourd'hui, quelques clics suffiraient à n'importe-qui pour transformer un fichier Word mis en page de manière autarcique en un fichier PDF, en annoncer la parution en quelques instants à des centaines de collègues et d'institutions dans une « mailing

²⁰ Précisons que la disparition de l'édition commerciale n'a rien à voir avec la survie de la forme imprimée qui, elle, relève d'une problématique tout à fait différente : il ne faut pas oublier que, si la transformation d'un texte imprimé en un outil informatisé performant est une opération longue et pénible, l'opération inverse ne soulève aucune difficulté. Sans doute, s'inspirant de la pratique dominante d'il y a quelques siècles, nos descendants les plus nostalgiques se rendront-ils chez le relieur pour faire assembler les centaines de pages virtuelles téléchargées en quelques secondes et à un coût pratiquement nul.

list » parfaitement ciblée, enregistrer le fichier sur son propre site ou celui de l'institution d'appartenance pour qu'il puisse être librement téléchargé, encaisser éventuellement par l'intermédiaire de Paypal les contributions volontaires destinées à l'amortissement des frais, ouvrir un forum de discussions destiné à accueillir critiques et suggestions, introduire des liens vers d'autres sites qui pourraient héberger d'éventuels comptes rendus. Pourquoi ne le fait-on pas encore massivement ? Paradoxalement, parce que la forme imprimée... requiert un investissement lourd, et cette nécessité représente une ligne de partage entre la littérature « blanche » — celle qui a été jugée digne d'un tel investissement — et la littérature « grise » ; celle qui, autrefois, se distinguait par la forme artisanale des caractères non proportionnels et la présence de pages sans justification à droite.

Mais ces obstacles ne sont pas les seuls. D'autres, en effet, relèvent à la fois, d'une part de la nature intrinsèque du Web et de son mode de fonctionnement ; d'autre part, de la politique des institutions culturelles en ce qui concerne le patrimoine dont elles ont la charge et les rapports qu'elles entretiennent avec le monde de l'enseignement et de la recherche.

Qu'est-ce que le Web ? Très vaste question à laquelle on ne peut certes essayer de répondre en quelques lignes et qui, comme tout le monde sait, peut recevoir quantité de réponses différentes, voire contradictoires. Mais « contradictoire » ne signifie pas « incompatible » : en fait, toutes les réponses possibles comportent une part de vérité, selon qu'on privilégie tel ou tel angle d'observation. Ainsi, le Web est sans conteste un espace de liberté, puisqu'il ouvre la voie à une expression sans entrave et à la mise en place de réseaux de communication à la fois multiples et rapides qui permettent de diffuser toute sorte d'informations, opinions, créations et, en retour, de prendre immédiatement connaissance de nouvelles et d'agissements qui, autrefois, seraient demeurés confidentiels. Il n'empêche qu'il s'agit également d'un outil d'intoxication qui ouvre la voie à toute sorte de propagande douteuse et permet de déguiser en vérités les rumeurs les plus invraisemblables ; de plus, il constitue aussi un outil de surveillance qui permet à une autorité politique ou à une entité commerciale de recueillir toute sorte d'informations d'ordre privé qui sont immédiatement utilisées ou pourraient l'être par la suite avec des mobiles peu rassurants. Le Web est un outil de connaissance, puisqu'on peut obtenir en l'espace d'un éclair des milliers de renseignements sur n'importe quel sujet ; mais c'est aussi un piège d'ignorance car, comme la bibliothèque de Babel jadis imaginée par Jorge Luis Borges, il contient, dans un mélange indissociable, tout ce qui est vrai, tout ce qui est presque vrai, tout ce qui est approximativement vrai ou faux, tout ce qui est faux.

La définition partielle que je retiendrais ici revient à considérer le Web comme une sorte de « foire de Babel ». La différence avec la bibliothèque de Babel est fondamentale : la bibliothèque imaginaire est secrète, ou du moins silencieuse et discrète, alors que la foire est une exposition bruyante et permanente, une course perpétuelle à la visibilité :

au début, il était bon d'être sur le Web pour avoir une longueur d'avance sur les concurrents ; maintenant, il est indispensable d'y être, et d'y être avec un battage suffisant pour pouvoir émerger de la masse. Dans un élan d'imagination, j'assimile parfois le Web à ce que pourrait être la grand-rue de Trantor, la capitale de l'empire galactique, au temps de sa splendeur la plus éclatante.

Le Web est donc l'une des réalisations les plus abouties de ce que des esprits clairvoyants ont appelé la « société du spectacle ». C'est un spectacle sur écran, à l'instar de la télévision. Mais dans le Web, le spectacle se veut interactif. Cela donne une fausse impression de puissance, alors que le parcours de l'internaute est en fait soigneusement encadré : dans la plupart des cas, il n'est qu'un « client », au sens très large du terme, auquel on ne montre que ce que l'on veut bien montrer, c'est-à-dire une série de chalands de toute nature²¹ ; et le client qui essaye de pénétrer dans le serveur, donc derrière la vitrine, est considéré à juste titre comme un pirate. C'est pourquoi il ne faut jamais oublier la double fonction de l'écran et la double connotation de ce terme : un écran, c'est certes quelque chose qui affiche, expose et invite au dialogue, mais en même temps c'est quelque chose qui sépare, filtre et cache.

Le Web est tout cela ; il est inutile de s'en plaindre, car il ne pouvait se développer que selon les lignes de force qui sont actuellement les siennes. Seulement, ce qui apparaît tout à fait compréhensible dans un mode de fonctionnement marchand, l'est beaucoup moins dans un contexte culturel ou scientifique. Or, dans presque toutes les institutions qui relèvent de ces domaines, l'utilisateur est traité malgré tout comme un client : il n'a pas accès aux données de base ;²² il ne peut que prendre connaissance des résultats d'une requête spécifique présentés de manière parcellaire et dont l'usage lui-même est expressément soumis à condition. Le moins que l'on puisse dire, c'est qu'il s'agit d'une conception extensive des droits de propriété intellectuelle et, dans ces conditions, il y a lieu de se demander si, dans certaines institutions publiques, la notion de « mise en valeur » n'est pas plus importante que celle de « service rendu ».

Ce point mérite quelques réflexions, et tant mieux si leur contenu représente le point de vue d'un vieux chercheur qui, tout en reconnaissant sa partialité, ne se considère tout de même pas comme un voleur à la fois potentiel et « potentiel » (c'est-à-dire comme

²¹ Ce terme est employé ici dans une acception qui n'implique pas nécessairement un but lucratif.

²² Exemple (tiré du site suisse *e-codices*) : « Toute modification, reproduction, publication, octroi de licence, vente d'une image, de métadonnées (descriptions des manuscrits) ou de tout autre information du *e-codices* sont formellement interdits. Pour des raisons techniques le téléchargement automatique de l'ensemble ou d'une partie du site web n'est pas autorisé. De telles tentatives sont enregistrées, sous réserve d'éventuelles poursuites ultérieures [c'est moi qui souligne]. Le téléchargement manuel de pages isolées est en revanche autorisé » (<<http://www.e-codices.unifr.ch/fr/info/terms>>). Le « siphonage » de la base est donc interdit, ce qui est le droit le plus strict de ses gestionnaires. Mais pourquoi prévoir des poursuites éventuelles si le téléchargement n'est impossible que pour des raisons techniques ? Dans un registre ironique, on pourrait plutôt imaginer que les « pirates » devraient être récompensés pour avoir su venir à bout d'un problème technique malencontreux.

un « gibier de potence »). Cette prise de position ne concerne, bien entendu, que les données de nature scientifique créées et rassemblées par un organisme public dans le cadre d'un programme de documentation dont le produit s'adresse à la communauté des chercheurs. Il s'agit, d'autre part, de considérations d'ordre moral qui dessinent les contours de ce qui serait souhaitable, tout en prenant acte de la situation telle qu'elle existe.

Lorsqu'une base de données²³ à finalité documentaire et destinée à une utilisation publique a été créée par un organisme public dans le cadre d'un programme financé par des fonds publics (ou privés à titre inconditionnel), il n'y a que quatre sortes d'opérations qui doivent être rigoureusement interdites pour des raisons déontologiques :

1. Le plagiat, soit l'appropriation pure et simple de la base et de sa paternité. Ce serait de la piraterie.
2. La substitution, soit l'appropriation de la gestion de la base, même s'il n'y a pas usurpation de paternité ; peu importe si le but de l'appropriation est lucratif ou non. Les humains ne sont pas des coucous.
3. La dénaturation, soit l'altération de tout ou partie du contenu de la base sans que la responsabilité des concepteurs soit expressément dérogée. Des modifications malencontreuses ou erronées ne doivent pas faire perdre la face à autrui.
4. L'oblitération, soit l'utilisation des données sans en citer la source. La paternité d'un travail, que celui-ci ait fait ou non l'objet d'une rémunération spécifique, doit toujours être reconnue et identifiée. Il faut rendre à César ce qui est de César.

En dehors de ces opérations interdites, la disponibilité sous une forme brute tabulaire des données et des résultats des requêtes devrait être entière et globale. On ne voit pas, en effet, en quoi et comment cette disponibilité lèserait les intérêts des concepteurs et des gestionnaires de la base, même lorsqu'un aspect important, sinon primordial, de l'existence de cette dernière réside dans la mise en valeur d'une institution ou d'un site. Toute crainte à ce sujet serait absurde, car il faut bien se rendre compte de ce que les données brutes ne serviraient à rien pour la grande majorité des utilisateurs. Pour les usages courants, il serait de toute manière plus avantageux pour tout le monde, car plus rapide et plus ergonomique, de se connecter au site en ayant recours à son interface et à son moteur de recherche.

Mais — pourrait-on alors objecter — à quoi bon disposer des données brutes ? Simplement, pour pouvoir les utiliser dans la constitution de nouvelles bases, aussi bien privées que publiques (dans ce dernier cas avec l'accord, bien entendu, des gestionnaires des bases sources).

Cette nécessité peut être illustrée par un exemple. Supposons qu'un historien de la culture écrite ait l'intention de procéder à une analyse bibliométrique de la production

²³ Ce terme désigne ici l'ensemble formé par un corpus de données structurées et les dispositifs permettant de l'interroger et de fournir des réponses à des utilisateurs.

d'incunables dans une aire géographique donnée. Ce dont il aurait besoin, c'est d'une liste complète des éditions imprimées dans ce territoire, comportant le nom des auteurs et les titres des œuvres, le lieu d'impression, la date, l'atelier typographique, le format et sans doute d'autres renseignements d'ordre matériel. Ces renseignements pourraient être puisés dans les deux répertoires actuellement disponibles sur le Web : l'*Incunabula Short Title Catalog* (ISTC) et le *Gesamtkatalog der Wiegendrucke* (GW). En revanche, l'historien devrait chercher ailleurs — ou plus vraisemblablement rassembler par ses propres moyens à partir d'informations très éparpillées — des données concernant les auteurs (tel que la date de mort, le statut social et professionnel) et les œuvres, afin d'élaborer une typologie permettant de suivre l'évolution des besoins et des goûts des lecteurs, ainsi que des politiques éditoriales mises en œuvre par les producteurs.

Il est évident que l'accès direct aux données brutes, lorsqu'elles existent, serait de nature à faciliter grandement la tâche de notre historien. Non seulement, mais les champs « typologie textuelle » de la base nouvellement créée pourraient enrichir en retour les répertoires existants : la possibilité de sélectionner d'un clic toutes les éditions de droit canon, ou de médecine, ou de grammaire, constituerait sans conteste une avancée non négligeable aux yeux de tout utilisateur.

La disponibilité des données brutes n'est pas un vain mot : sur les sites gouvernementaux des États-Unis — pays où l'économie marchande jouit pourtant des plus grandes faveurs — il est possible de télécharger sans problème des fichiers « texte » contenant les coordonnées géographiques de tous les lieux peuplés du monde ou l'historique des observations météorologiques pour toutes les villes américaines, et de les utiliser sans restriction.²⁴

Quelles sont les perspectives d'avenir ?

Dans le domaine de la culture écrite, le cas de l'incunable peut être envisagé avec un certain optimisme. La nature même de l'objet interdit le repli sur soi : il ne nous reste, en effet, qu'environ 28000 éditions, et la description établie à partir d'un seul exemplaire peut s'appliquer automatiquement à tous les autres, sauf, bien entendu, pour ce qui relève du chemin parcouru par chaque exemplaire après parution.²⁵ C'est pour cette raison que le répertoriage total des éditions a pu être entrepris depuis longtemps, que le recensement exhaustif des exemplaires survivants a été achevé dans la presque totalité des pays et que les grandes bibliothèques de conservation ont pu établir le catalogue de leur fonds dans des délais raisonnables.²⁶ Dans ce contexte, le processus

²⁴ C'est le cas pour la *National Geospatial Intelligence Agency* (NGA) qui gère le *GEOnet Names Server* (GNS) : « Geographic names data is freely available. A suitable citation note is : "Toponymic information is based on Geographic Names Database, containing official standard names approved by the United States Board on Geographic Names and maintained by the National Geospatial-Intelligence Agency ».

²⁵ Il est vrai, aussi, qu'il n'existe pas d'édition dont tous les exemplaires sont parfaitement identiques, mais les différences sont le plus souvent négligeables.

²⁶ Celui de la *Bayerische Staatsbibliothek* de Munich a été mis en ligne. La numérisation des éditions est en cours.

de numérisation, déjà bien entamé en Allemagne²⁷, peut s'appuyer sur des principes rationnels : le fait que les livres les plus remarquables sur le plan esthétique sont ceux qui ont été conservés en plus grande quantité interdit de privilégier les « trésors », et la numérisation prioritaire des éditions autochtones ou de celles qui contiennent des ouvrages écrits dans la langue nationale n'apparaît plus comme l'expression de velléités identitaires extrascientifiques, mais comme un partage du travail dicté par des critères somme toute cohérents, puisqu'il est à peu près certain que chaque pays assumera sa part du fardeau.

Pour ce qui est du manuscrit, en revanche, rien n'incite à être résolument optimiste. À vrai dire — mais c'est peut-être une impression personnelle — je constate que la plupart des institutions, surtout les plus grandes et prestigieuses, tendent à jouer en solo ; leur mode de fonctionnement me paraît singulièrement « monadique », au sens leibnizien du terme : elles ne veulent pas se dessaisir de leurs précieuses données, même à titre provisoire, même pour permettre la constitution d'ensembles harmonisés et intégrés pouvant être interrogés simultanément.²⁸ Dès lors, il faut s'attendre à ce que ces comportements monadiques produisent une série de chants monodiques, sans doute dissonants, et tout ce que l'on peut espérer, c'est que la recherche d'un minimum d'harmonie ne se fera pas au détriment des aspirations légitimes des chercheurs.

Bibliographie

- Baryla, Christiane. « La banque d'images des manuscrits de la Bibliothèque Vaticane. I. La réalisation des vidéo-disques ». *Gazette du livre médiéval* 20 (1992) : 16–26.
- Baschet, Jérôme. « La banque d'images des manuscrits de la Bibliothèque Vaticane II. L'indexation iconographique ». *Gazette du livre médiéval* 20 (1992) : 26–29.
- Bernardi, Francesco, Paolo Eleuteri et Barbara Vanin. « La catalogazione in rete dei manoscritti delle biblioteche venete : Nuova Biblioteca Manoscritta ». *KPDZ* 1. 3–11.
- Brepolis. *The hometown of Brepols' online publications*. Turnhout : Brepols Publishers, 2010. <<http://www.brepolis.net/>>.
- British Library *Catalogue of Illuminated Manuscripts*. Londres : British Library, 2010. <<http://www.bl.uk/catalogues/illuminatedmanuscripts/welcome.htm>>.

²⁷ Cette entreprise (la »Verteilte digitale Inkundebibliothek« ou »VDLB«) est gérée par la Herzog August Bibliothek (HAB) de Wolfenbüttel, ainsi que par la Universitäts- und Stadtbibliothek (USB) de Cologne.

²⁸ A côté de leurs indéniables mérites, il convient par ailleurs de souligner les incontestables limites des opérations visant à l'intégration sur le Web de bases de données ayant une thématique commune, ce qui est tout de même autre chose que la simple normalisation, grâce à une métabase conçue *ad hoc*, des noms d'auteurs et des titres d'œuvres. Bien souvent, la nécessité impose de se contenter d'un simple « coup de peinture » qui aboutit à la création d'un portail permettant une interrogation simultanée des bases intégrées. Il ne s'agit en aucun cas d'une véritable harmonisation, car on ne modifie pas le contenu, et encore moins les principes — souvent fort différents — qui ont présidé à la conception de chaque base. Dans ces conditions, l'interrogation simultanée se restreint aux seuls champs, en général peu nombreux, communs à toutes les bases, à condition que leur contenu ait préalablement été uniformisé.

- Cartelli, Antonio et al. « Il catalogo aperto dei manoscritti Malatestiani ». *KPDZ* 1. 13–23.
Catalogo aperto dei manoscritti Malatestiani. Cesena : Biblioteca Malatestiana, 2002–2010.
<http://www.malatestiana.it/manoscritti>.
- Collectif de rédaction. « Déclaration des droits du manuscrit, du lecteur et du conservateur ». *Gazette du livre médiéval* 14 (1989) : 1–4.
- Consortium of European Research Libraries (CERL) *Thesaurus*. Londres : CERL, 2007–2010.
http://www.cerl.org/web/en/resources/cerl_thesaurus/main.
- Degl'Innocenti, Emiliano. « Il Progetto di digitalizzazione dei Plutei della Biblioteca Medicea Laurenziana di Firenze ». *DigItalia* 1 (2007) : 103–113.
http://digitalia.sbn.it/upload/documenti/digitalia20071_DEGLINNOCENTI.pdf.
- Digitale Handschriften. Large Digital Libraries of Western Manuscripts*. Éd. par Klaus Graf. Lüneburg : Netbib.de, 2009–2010. <<http://wiki.netbib.de/coma/DigitaleHandschriften>>.
- Digitalizzazione dei manoscritti del fondo antico presso il Sacro convento di Assisi*. Assisi : Società Internazionale di Studi Francescani. <<http://www.sisf-assisi.it/digitalizzazione.htm>>.
- e-codices. Bibliothèque virtuelle des manuscrits en Suisse*. Fribourg : Université de Fribourg, 2008–2010. <<http://www.e-codices.unifr.ch/fr>>.
- GEOnet Names Server (GNS)*. Washington : National Geospatial Intelligence Agency (NGA). <<http://earth-info.nga.mil/gns/html/index.html>>.
- Gesamtkatalog der Wiegendrucke (GW)*. Berlin : Staatsbibliothek Preussischer Kulturbesitz. 2007–2010. <<http://www.gesamtkatalogderwiegendrucke.de/>>.
- Incunabula Short Title Catalog (ISTC)*. Londres : British Library, 1980–2010.
<http://www.bl.uk/catalogues/istc/index.html>.
- Inkunabelkatalog der Bayerischen Staatsbibliothek (BSB-Ink online)*. Munich : Bayerische Staatsbibliothek, 2010. <<http://www.bsb-muenchen.de/Inkunabeln.181.0.html>>.
- Digitale Sammlungen* : <<http://www.bsb-muenchen.de/Digitale-Sammlungen.72.0.html>>.
- KPDZ 1 : Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Éd. par Malte Rehbein, Patrick Sahle et Torsten Schaßan. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt : Books on Demand, 2009. En ligne : <urn:nbn:de:hbz:38-29393>, <<http://kups.ub.uni-koeln.de/volltexte/2009/2939/>>.
- Manuscripta Mediaevalia*. Éd. par Zentralredaktion mittelalterlicher Handschriftenkataloge. Berlin, Marburg, München [1996–2010] <<http://www.manuscripta-mediaevalia.de/>>.
- Mirabile. Archivio digitale della cultura latina medievale*. Firenze : Società internazionale per lo studio del Medioevo latino (SISMEL) – Fondazione Ezio Franceschini (FEF).
<http://www.mirabileweb.it/>.
- Nuova Biblioteca Manoscritta (NBM)*. Venise : Regione del Veneto et Università Ca' Foscari di Venezia, 2006–2010. <<http://www.nuovabibliotecamanoscritta.it/>>.
- Ornato, Ezio. « 'Bibliotheca manuscripta universalis'. Digitalizzazione e catalografia : un viaggio nel regno di Utopia ? ». *Gazette du livre médiéval* 48 (2006) : 1–13.
- Progetto Irnerio*. Bologna : Università di Bologna, Centro interdisciplinare in storia e filosofia del diritto e informatica giuridica (CIRSFID). <<http://irnerio.cirsfid.unibo.it/>>
- Rizzo, Silvia. « Conservation et jouissance du patrimoine manuscrit : quelques réflexions ». *Gazette du livre médiéval* 4 (1984) : 14–17.

Smith, Marc. « Fac-similés de manuscrits ». *Ménestrel*, 19.01.2010.

<<http://www.menestrel.fr/spip.php?rubrique731>>.

Teca digitale. Florence : Biblioteca Medicea Laurenziana.

<<http://teca.bmlonline.it/TecaRicerca/index.html>>.

Text Encoding Initiative (TEI). TEI Consortium, 2010. <<http://www.tei-c.org/>>. Manuscript Description : <<http://www.tei-c.org/release/doc/tei-p5-doc/html/MS.html>>.

Union List of Artist Names Online (ULAN). Los Angeles : J. Paul Getty Museum.

<http://www.getty.edu/research/conducting_research/vocabularies/ulan/>.

Verteilte Digitale Inkunabelbibliothek (VDLB). Wolfenbüttel : Herzog August Bibliothek – Cologne : Universitäts- und Stadtbibliothek Köln. <<http://inkunabeln.ub.uni-koeln.de/>>.

Vitale Brovarone, Alessandro. « Lector cavat codicem ? ». *Gazette du livre médiéval* 6 (1985) : 13–16.

Applying Semantic Web Technologies to Medieval Manuscript Research

Toby Burrows

Abstract

Medieval manuscript research is a complex, fragmented, multilingual field of knowledge, which is difficult to navigate, analyse and exploit. Though printed sources are still of great importance and value to researchers, there are now many services on the Web, some commercial and many in the public domain. At present, these services have to be consulted separately and individually. They employ a range of different descriptive standards and vocabularies, and use a variety of technologies to make their information available on the Web. This chapter proposes a new approach to organizing the international collaborative infrastructure for interlinking knowledge and research about medieval European manuscripts, based on technologies associated with the Semantic Web and the Linked Data movement. This collaborative infrastructure will be an open space on the Web where information about medieval manuscripts can be shared, stored, exchanged and updated for research purposes. It will be possible to ask large-scale research questions across the virtual global manuscript collection, in a quicker and more effective way than has ever been feasible in the past. The proposed infrastructure will focus on building links between data and will provide the basis for new kinds of services which exploit these data. It will not aim to impose a single metadata standard on existing manuscript services, but will build on existing databases and vocabularies. The article describes the architecture, services and data which will comprise this infrastructure, and discusses strategies for making this challenging and exciting goal a reality.

Zusammenfassung

Die Handschriftenforschung ist ein umfangreiches, fragmentiertes und multilinguales Wissensgebiet, das nicht leicht zu überblicken, zu analysieren oder auszuschöpfen ist. Obwohl gedruckte Quellen für die Forschung weiterhin von großer Bedeutung und Nutzen sind, gibt es inzwischen viele Angebote im Web, teils kommerzieller Natur, teils frei zugänglich. Im Augenblick können diese Angebote jedoch nur isoliert voneinander in Anspruch genommen werden. Sie fußen auf einer Reihe verschiedener Beschreibungsstandards und Vokabularien und nutzen eine Bandbreite

unterschiedlicher Technologien, um ihre Informationen im Web zur Verfügung zu stellen. Dieser Beitrag beschreibt einen neuen Ansatz, um Wissen und Forschung über mittelalterliche europäische Handschriften zusammenzuführen: eine internationale, kollaborative Infrastruktur, die auf den Technologien aus den Bereichen “Semantic Web” und “Linked Data” aufbaut. Daraus entsteht im Web ein “open space”, in dem Informationen zu mittelalterlichen Handschriften gespeichert, bereitgestellt, ausgetauscht und aktualisiert werden können. Damit wird systematische Forschung auf Grundlage einer globalen, virtuellen Handschriftensammlung schneller und zielführender als bislang denkbar. Die vorgeschlagene Infrastruktur konzentriert sich auf die Verknüpfung von Daten, und sie wird die Grundlage für neuartige Angebote bilden, um mit diesen Daten zu arbeiten. Sie zielt nicht darauf, einen bestimmten Standard für Metadaten durchzusetzen, sondern baut auf den bestehenden Datenbanken und Vokabularen auf. Der Beitrag beschreibt die Infrastruktur im Hinblick auf Systemaufbau, Services und Daten und erörtert Strategien zu ihrer Realisierung.

1. Introduction

Among the greatest treasures in the cultural heritage of Europe are its medieval manuscripts. Many hundreds of thousands still survive today, in collections around the world, where they are intensively studied by researchers and admired by visitors to libraries, museums and art galleries. There are numerous Web sites and projects devoted to medieval manuscripts, originating both from research groups working on these manuscripts and from the cultural heritage institutions in which they are held. In many ways, medievalists have been at the forefront of the application of digital technologies to research in the humanities.

Despite their undoubted value, these Web sites and services suffer from several major limitations. There is a lack of integration and interoperability between the many different sites, and it is difficult to find out systematically what research and digitization are being undertaken in collections around the world. Most of these services have to be consulted separately and individually, though search engines like Google cover some of them. The use of terminology and of descriptive standards is inconsistent and unsatisfactory, especially across different European languages and cultures. There is often no easy way of connecting the descriptions of the manuscripts with the reports of the research which has been based on them. As a result, researchers around the world still face major difficulties in finding, using and sharing knowledge about medieval manuscript collections.

This paper presents a proposal for a new international collaborative infrastructure for organizing and interlinking knowledge and research about medieval European manuscripts, based on technologies associated with the Semantic Web and the

Linked Open Data movement. The proposal draws on a Road Map developed for a European Science Foundation Exploratory Workshop held at the University of Birmingham (United Kingdom) in March 2009. Convened by Wendy Scase (University of Birmingham) and Orietta Da Rold (University of Leicester), this Workshop brought together specialists from the fields of manuscript studies, librarianship, and information and computer science, as well as from a range of public and commercial institutions and organizations.

The Exploratory Workshop was organized by the Medieval Manuscript Research Group of CARMEN, the Co-operative for the Advancement of Research through a Medieval European Network. CARMEN includes representatives of the many European centres for medieval research, as well as professional associations, cultural institutions and commercial publishing companies with expertise in this field. Participants from outside Europe include the Australian Research Council's Network for Early European Research (NEER) and several North American associations and institutions. A large proportion of the global research community in this field (estimated at more than 14,000 researchers) is represented in CARMEN.

2. Medieval Manuscripts: Research Questions

Medieval manuscripts are used in addressing a wide range of research questions. Most obviously, these include **research into the characteristics of manuscripts themselves as physical objects**. These characteristics include: the place of origin, the date or period of origin, the materials used, the decoration and illumination, the handwriting, the scribe, the binding, arrangement of the physical volume, and the language. Research into the subsequent history of a manuscript looks at its owners and at changes to its appearance over time, as well as at its modern location, and its place in modern collections.

Relationships between manuscripts are a common topic, including research which reunites dispersed leaves of what was originally a single manuscript. Identifying connections between specific medieval manuscripts and other materials which survive from the medieval period (especially art works, buildings, and other material objects) is another closely related area of research.

Defining these different physical characteristics also forms the starting-point of many research projects. These include such investigations as defining specific styles of handwriting, establishing different categories of decoration, and identifying different ways of constructing and creating physical volumes in the medieval period.

The second major area of research involves **the use of manuscripts as evidence for all aspects of life in the medieval period**. This requires knowledge and understanding of the *content* of a manuscript—the text, the images, the music, and so on. This kind of

research is heavily dependent on the descriptors used to identify the content, including authors' names, titles of works, incipits, subject and concept terms, and so on.

Both of these two major areas of research also draw on an extensive body of secondary literature relating to specific manuscripts: catalogues and descriptions (both medieval and modern), bibliographies, and secondary works. These are likely to reflect changes over time—as concepts shift, and descriptions and attributions are revised. All aspects of the body of knowledge in this field are multilingual; descriptions and descriptors may be in a variety of languages, mainly European.

Fundamental to both types of research projects is the availability of detailed descriptions of manuscripts as physical objects. In the world of printed catalogues, there are numerous different formats for such descriptions. In the digital world, a similar multiplicity of formats was already evident as long ago as the late 1980s and early 1990s. The major change since then has been the emergence of the Manuscript Description element (<msDesc>) and chapter (10. Manuscript Description) within the encoding guidelines of the Text Encoding Initiative (TEI), which is being increasingly regarded as a de facto standard. The North American Association of College and Research Libraries has also published rules for the *Descriptive Cataloging of Ancient, Medieval, Renaissance, and Early Modern Manuscripts (Pass)*.

3. Web Resources for Medieval Manuscript Research

Medieval manuscript research is a complex, fragmented, multilingual field of knowledge, which is difficult to navigate, analyse and exploit. Though printed sources are still of great importance and value to researchers, there are now many services on the Web, some commercial and many in the public domain. The number of such services is hard to estimate, but listings of the major authoritative sites (such as those in the Labyrinth and the Virtual Library) contain at least fifty entries. These services employ a range of different descriptive standards and vocabularies, and use a variety of different technologies to make their information available on the Web.

Numerous collecting institutions provide information about the manuscripts they hold, either as part of more general databases or as specific manuscript databases. Examples of the latter approach include the British Library's Manuscripts Catalogue and the Codices Electronici Sangallenses (CESG) of the Abbey Library of St Gall. There are also a range of national databases, including Medieval Manuscripts in Dutch Collections (hosted by the Koninklijke Bibliotheek), MEDIUM (hosted by the Institut de recherche et d'histoire des textes), and the Digital Scriptorium, hosted by Columbia University and the University of California Berkeley. The Europa Inventa project has developed a national database for Australia.

The small number of international databases are particularly important in the context of this paper. The CERL (Council of European Research Libraries) Portal provides a union catalogue of manuscript descriptions, harvested from sites distributed across Europe, North America and Australia. The Manuscriptorium service, hosted by the Národní knihovna České republiky, contains descriptive records contributed by more than ninety institutions across Europe.

Some of these services provide digital images of some manuscripts as well as descriptive information about them. Others focus specifically on providing digital versions. The European digital library service Europeana, for instance, only includes digitized materials, though its scope is much broader than medieval manuscripts. Its content is being enhanced by a new project, Europeana Regia, which aims to digitize a corpus of illuminated royal manuscripts from France, Belgium, Germany and Spain. Large-scale digitization projects are underway in several major libraries, including the Bayerische Staatsbibliothek and the Biblioteca Apostolica Vaticana.

There are many Web sites which list, transcribe, or provide digital images of manuscripts of a specific text or relating to a specific medieval author. Two examples are the site devoted to Dante's *Divina Commedia*, maintained by the Società Dantesca Italiana, and the interrelated group of sites focusing on *Le Chevalier de la Charrette* by Chrétien de Troyes, hosted by Princeton University, Baylor University and the Université de Poitiers.

Ancillary Web services include sites devoted to manuscript terminology and vocabularies, incipits, subjects, authors, and people more generally. Some of these sites are commercial, notably the various databases managed by Brepols Publishers NV in Belgium. These include In Principio (incipits), the International Medieval Bibliography (IMB) with its subject thesaurus and lists of authors, and Europa Sacra (medieval prelates). Many other Web resources of this kind are hosted and maintained by cultural institutions. The CERL Thesaurus is an extensive collection of place names and personal names. The Medieval Manuscripts in Dutch Collections service provides its own list of authors. Personennamen des Mittelalters is hosted by the Deutsche Nationalbibliothek as part of its larger database of personal names (Personennamendatei). The *Fasti Ecclesiae Anglicanae 1066–1300* focuses on English church prelates. Also important is the Web version of Denis Muzerelle's *Vocabulaire codicologique*, hosted by the IRHT. In Italy, the SISMEL research group publishes two major databases of medieval authors' names through its Mirabile service: *Bibliotheca Scriptorum Latinorum Medii Recentiorisque Aevi* (BISLAM) and *Compendium Auctorum Latinorum Medii Aevi* (CALMA).

Other services provide indexes to references to specific manuscripts in journal articles, scholarly books and other secondary literature. They include the indexes to the journal *Scriptorium* and the manuscripts index within the International Medieval Bibliography.

This extraordinary—and continually growing—collection of riches brings its own set of complications and difficulties, and leaves us with something resembling the Tower of Babel. There are simply too many sources of information, with manuscript descriptions in different formats and multiple languages, and highly variable uses of names, titles and concepts. It is very difficult to trace systematically the relationships between different manifestations of the same manuscript (e.g., digital images, transcriptions and editions). It can also be very difficult to trace relationships between manuscripts and what is written about them (e.g., articles, books and commentaries).

While the proliferation of Web resources is undoubtedly of great value to manuscript research, their sheer complexity and variety impose a significant barrier to our ability to ask large-scale research questions—both about manuscripts as physical objects and about their content. Trying to impose rigorous standards and continuing to pursue uniformity, in order to resolve these difficulties, is unlikely to be successful. Expecting Google—or some future global search engine—to provide a solution through keyword searching is equally unrealistic. A different kind of approach is required.

4. Semantic Web Technologies and the Linked Data Movement

If a serious effort is to be made to overcome the complexities of the current digital landscape, the most promising approach appears to lie in what are somewhat misleadingly called Semantic Web technologies. The underlying idea of the Semantic Web is to implement methods of making digital content available for processing by Web-based software in a more effective way, though its originator, Tim Berners-Lee, now prefers to use the term “Web of data”, to emphasize the data-oriented nature of this framework. The primary purpose of these developments is to enable knowledge to be found, shared and interlinked more easily.

Semantic Web technologies are methods for adding semantic structures to Web data and documents, with the broad aim of making them more interoperable and automatically discoverable (Antoniou and van Harmelen). The main building-blocks for this process are as follows:

- **Object identifiers:** alpha-numeric addresses such as URIs (Uniform Resource Identifiers) which can be used to identify an object uniquely. It is possible to assign identifiers to abstract “objects” like concepts, subject terms, personal names and place names, as well as to physical objects like manuscripts.
- **Ontologies and ontological languages:** ways of structuring the relationships within the vocabulary and terminology of a body of knowledge, expressed in a formal language like OWL (Web Ontology Language) or SKOS (Simple Knowledge Organization System). The result is a machine-readable conceptual map of a domain of knowledge.

- **RDF databases:** collections of statements about objects, their properties, and their relationships (“triples”), expressed in the RDF (Resource Description Framework) syntax. These statements can be used to show how and where an object fits in the ontological structure of the body of knowledge.
- **Agent systems and Web services:** software environments which can be built to explore, analyse and exploit the knowledge embedded in ontologies and RDF databases.

Several major European research projects have already been applying Semantic Web technologies to the cultural heritage domain, including MultimediaN in the Netherlands (Schreiber et al.) and CultureSampo in Finland (Hyvönen et al.). MultimediaN, in particular, has been used to provide the prototype for the Europeana service. But these projects are still aimed at developing standalone Web services—albeit of a particularly sophisticated kind—and in some ways might be seen as adding to the complexity of the digital landscape rather than enabling researchers to make more effective use of it.

Semantic Web technologies are also being employed to develop a different kind of service, which aims to make data available on the Web in formats which can be processed automatically. A rapidly growing number of datasets are now available under the umbrella of the Linked Data movement (Bizer et al.). As of April 2010 they already included the Personennamendatei of the Deutsche Nationalbibliothek, as well as authoritative library vocabularies like the Library of Congress Subject Headings and RAMEAU (Bibliothèque nationale de France). These services simply provide the data; other services which enable the navigation and use of the whole body of linked data are developed separately and independently.

These technologies are intended to represent a complex body of knowledge in a connected and interoperable way, and to provide a platform for applying sophisticated discovery and analytical tools to data across the Web. They are likely to be of particular value for medieval manuscript research in the following ways:

- improving interoperability and interconnection between the many manuscript-related Web sites;
- interlinking the variety of different terminologies, vocabularies and data standards relating to manuscripts, particularly in the European multilingual and multicultural environment;
- enabling more direct connections between the manuscript descriptions and catalogues produced by cultural institutions and the continually developing apparatus of scholarly annotation, editing, study and commentary derived from and based on those manuscripts.

5. Linked Data for Medieval Manuscript Research

Semantic Web technologies and the Linked Data movement can be harnessed to build a new international collaborative infrastructure for organizing and interlinking knowledge and research about medieval European manuscripts. This collaborative infrastructure will be an open space on the Web where data about medieval manuscripts can be shared, stored, and exchanged for research purposes. It will focus on building links between data and will provide the basis for developing new kinds of services which exploit these data.

On the other hand, this infrastructure will not encroach on areas where cultural institutions exercise intellectual property rights over the medieval manuscripts in their custody, such as image reproduction. Nor should it aim to impose a single metadata standard on existing manuscript-related services; instead it should join up existing databases and vocabularies, such as those in services like the CERL Portal and Manuscriptorium. The proposed collaborative infrastructure is not intended to replace these existing databases and catalogues, but instead to connect them and act as a broker between them.

The proposed infrastructure will not, of itself, form a “virtual research environment”. It will not include a transcription or digitization service, nor will it consist of a TextGrid for the analysis and manipulation of medieval texts. Nevertheless, such services will be able to link into the proposed infrastructure and to use its features—such as its manuscript identifiers—as points of reference. It will serve as the framework on which virtual research environments can be built. It is a necessary first stage before global services for asking research questions across different databases and different formats (digital, print, and manuscript) can be constructed.

Making this first stage a reality will require the development of a Linked Data environment for medieval manuscript research. This will involve transforming existing knowledge into Semantic Web formats and making it available on the Web. The main source of this knowledge is the extensive semantic network already embedded in existing manuscript descriptions, covering such semantic categories as names, identifiers, quantities, concepts, manifestations and dependents. Some existing services focus on just one of these categories, while others—particularly manuscript databases and library catalogues—cover a wider range.

Assigning unique identifiers to manuscripts is the key starting-point.¹ A URI for each manuscript will provide the crucial reference point which can be used as the basis for linking other kinds of information about that manuscript. A service for resolving these identifiers to gain access to relevant descriptions and related objects will be an essential component. Identifier services will also need to cover names, places, events and other

¹ I thank Antoine Isaac for his valuable contribution to the analysis of the requirements for this framework.

Semantic categories	Data types	Existing sources and services
IDENTIFIERS	shelf numbers, catalogue numbers, reference numbers	used in manuscript databases and library catalogues, printed catalogues, IMB, Scriptorium
NAMES	people (authors, owners, editors, bibliographers, commentators, artists), institutions, places	Europa Sacra, IMB, CERL Thesaurus, Personennamen des Mittelalters
CONTENTS	titles of works, incipits	In Principio, IMB, Scriptorium, manuscript databases and library catalogues, printed catalogues
QUANTITIES	measurements, sizes	recorded in manuscripts databases, library catalogues, and in printed catalogues
CONCEPTS	subject matter, materials, scripts, type of work, languages	IMB, Getty vocabularies, Icon-Class
EVENTS	dates and times	IMB, manuscripts databases and library catalogues, printed catalogues
MANIFESTATIONS	editions, transcriptions, images (printed and digital)	Web sites and books; also listed in IMB, library catalogues, some manuscript databases, printed catalogues
DEPENDENTS	scholarly writings, annotations, commentaries, bibliographical entries	listed in IMB, Scriptorium, some manuscript databases

Table 1. Semantic categories in medieval manuscript descriptions.

conceptual structures and entities. Some of these have already been developed by other disciplines, and can be reused or adapted.

Vocabularies will be another crucial part of the service, formed either by transforming existing vocabularies into an appropriate format like SKOS (Isaac and Summers) or by extracting terms from descriptive database records through techniques like text mining. Building alignments between these vocabularies is an important requirement; this will enable different vocabularies to be mapped and the relationships between terms to be indicated, without necessarily having to select one term as more authoritative than another. This will be of particular value in dealing with such areas as scripts and handwriting, where there is no generally accepted vocabulary and where the terminology used varies from country to country. Mapping between different languages will also be a significant issue here.

Closely related is the requirement to develop schemas for descriptive structures—in other words, identifying the different components of manuscript descriptions and mapping their variations. Building alignments and mapping between different languages will also be important in this area. These schemas—together with the vocabularies and their alignments—will need to be stored in repositories which are capable of providing terminology services. In practice, this will require a RDF “triple store” or similar technology (Hertel et al.).

Inherent in this framework will be a graph (in the sense of an abstract data structure intended to implement mathematical graph theory) showing the many different relationships within the data. These relationships will include those between different vocabularies, between different entities, between the component elements of a schema, between manuscripts and the entities which describe them, between manuscripts and their different manifestations, between manuscripts and their various dependents, and so on. Maintaining these interoperable knowledge bases for constructing manuscript descriptions will be an important challenge; both automated and manual methods will need to be tested and applied.

Once these data stores are in place, it will become possible to build Web services which exploit them and which add value to our existing knowledge about manuscripts. These should include services for querying and browsing the data, as well as visualizations which draw on the relationships inherent in the data structures, and map-based interfaces derived from the geographic information included in the data stores. Interfaces which enable annotation and comments from the scholarly community are another possible layer, together with services which allow the semantic network to be updated—either manually or in an automated way.

6. Organizational Requirements

The scale and complexity of the proposed infrastructure mean that it is unlikely to be developed quickly or as the result of a single, centralized project. A more feasible approach would involve making a start with a small number of datasets—and possibly several parallel projects—within an agreed overall technical framework. Existing services and institutions can be encouraged to make their data available for reformatting and inclusion.

Research groups and individual researchers need to be involved, in order to contribute to the analysis of use cases and user requirements, and to ensure that the resulting products and services are relevant to their needs. They will also be required for testing, correcting and updating, particularly once annotation and commenting become available. Their involvement can be harnessed through organizations like CARMEN, as well as directly through suitable interfaces to the data stores. The interest and

enthusiasm of the postgraduate and postdoctoral student community can also be drawn on, through training networks and other capacity-building initiatives.

Libraries and other collecting institutions are of crucial importance. As well as being the custodians of most of the manuscripts, they continue to make a major contribution by constructing and maintaining descriptive databases and managing digitization programmes and services. Their involvement can be harnessed through existing co-operative arrangements for cataloguing and digitization and through organizations like CERL, as well as individually. Commercial firms can also make a significant contribution; publishing houses specializing in medieval studies will have data and subject knowledge to contribute, while technology companies can contribute their technical expertise.

The proposed infrastructure provides an excellent opportunity for the public and private sectors to work together through the pooling of mutually relevant data and technological expertise. It will also work to strengthen links between academic researchers and the curatorial and professional staff in cultural institutions (libraries, museums, galleries and archives). Both groups have a vital and enduring interest in medieval manuscripts, but their priorities and perspectives are significantly different. Developing a Linked Data infrastructure of this kind will encourage them to share and pool their knowledge and expertise.

It should be possible to learn from other disciplines—especially in the sciences—which are already actively building global knowledge bases. This is not to say that the model based mainly on a single institution or on a loose coalition of institutions, which is probably more characteristic of the humanities, is completely inappropriate. The Perseus project, for example, has assembled a remarkably successful digital library service in classical studies from a base at Tufts University. But the scale involved in building a global infrastructure for manuscript data is more likely to require a globally coordinated approach.

An interesting model is provided by the Shared Names project, associated with the Science Commons and NeuroCommons collaborative initiatives. This project aims to assign URIs as names for biomedical information records (primarily genes), using a community-managed shared infrastructure for maintenance and development. Also of relevance from the biomedical sciences is the Encyclopedia of Life. Though not a Semantic Web project, it uses an innovative administrative structure which draws on the contributions of researchers and individuals around the world. These and other similar projects suggest that decentralized and distributed organizational arrangements which encourage individual contributions may be the most effective way of approaching a global initiative of this kind.

7. Conclusion

The surviving manuscripts are one of the key sources for research into medieval Europe and are used in addressing a wide range of research questions. Most obviously, these include research into the characteristics of manuscripts themselves as physical objects. The other major focus of research involves the use of manuscripts as evidence for all aspects of life in the medieval period.

Medieval manuscript research is a complex, fragmented, multilingual field of knowledge, which is difficult to navigate, analyse and exploit. Though printed sources are still of great importance and value, there is a large and rapidly growing body of material on the Web. Much of this Web material consists of descriptive information about manuscripts, though a considerable amount of digitization and transcription has also been carried out. At present, however, the digital landscape is difficult to navigate and overwhelming in its richness and complexity.

The infrastructure proposed in this paper focuses on the possibilities for applying new Semantic Web technologies to medieval manuscript research. These technologies have the potential to enhance research greatly, by enabling much more effective access to, and use of, relevant materials and knowledge. It will be possible to ask large-scale research questions across the global manuscript collection, in a quicker and more effective way than has ever been feasible in the past. The ultimate goal should be a Web service through which a researcher can readily find all manuscripts of relevance to the research question she is investigating, and be pointed to previous work about them and to digital representations of them.

The technologies which can make this happen are now coming to maturity, though it will ultimately require a global effort to harness their full potential. But we are in a position to envisage what the “Web of data” for medieval manuscript research might look like, and to start organizing ourselves to make it a reality.

Bibliography

- Angjeli, Anila, et al. “Semantic Web and vocabulary interoperability: An experiment with illumination collections.” *International Cataloguing and Bibliographic Control*, 38.2 (2009): 25–29.
- Antoniou, Grigoris, and Frank van Harmelen. *A Semantic Web Primer*. 2nd ed. Cambridge, MA: MIT Press, 2008.
- Berners-Lee, Tim. “The next Web of open, linked data.” 2009.
http://www.youtube.com/watch?v=OM6XIICm_qo
- Bizer, Christian, Tom Heath and Tim Berners-Lee. “Linked Data – The Story So Far.” *International Journal on Semantic Web and Information Systems* (2009).
<http://linkeddata.org/docs/ijswis-special-issue>.

- British Library's Manuscripts Catalogue*. [Search entry to the main catalogues of the British Library's collection of Western manuscripts.] London: British Library, 2002–2005.
<<http://www.bl.uk/catalogues/manuscripts/INDEX.asp>>.
- CARMEN. *Co-operative for the Advancement of Research through a Medieval European Network*. Groningen: University of Groningen, Research School Medieval Studies 2003–2007.
<<http://www.carmen-medieval.eu/>>.
- Carusi, Annamaria and Torsten Reimer. "Text-Grid Virtual Research Environment in the e-Humanities." *Virtual research environment collaborative landscape study – A JISC funded project*. January 2010: 92–94.
<<http://www.jisc.ac.uk/media/documents/publications/vrelandscapeproject.pdf>>.
- The CERL Portal: Manuscripts and Early Printed Material*. Uppsala: Consortium of European Research Libraries, 2006–2007. <<http://cerl.epc.uu.se/portal>>.
- CERL Thesaurus*. Göttingen: Consortium of European Research Libraries, 2008
<<http://thesaurus.cerl.org/>>.
- [Chrétien de Troyes, *Le Chevalier de la Charrette*.] The Charrette Project 2. *Princeton University, Baylor University, and Université de Poitiers*, 2005. <<http://lancelot.baylor.edu/>>.
- CESSG: Codices Electronici Sangallenses*. University of Fribourg and Abbey Library of St. Gall, 2005–2010. <<http://www.cesg.unifr.ch/en>>.
- Crane, Gregory., Brent Seales and Melissa Terras. (2009). "Cyberinfrastructure for classical philology." *Digital Humanities Quarterly*, 3.1 (2009).
<<http://www.digitalhumanities.org/dhq/vol/3/1/000023.html>>.
- [Dante Alighieri, *La Divina Commedia*. Index of Manuscripts] Dante online. Indice dei Manoscritti. Società Dantesca Italiana, 2002.
<http://www.danteonline.it/italiano/codici_indice.htm>.
- Digital Scriptorium*. New York: Columbia University Libraries, 2007.
<<http://www.scriptorium.columbia.edu/>>.
- Encyclopedia of Life*. Washington D.C.: Smithsonian Institution et al., 2010.
<<http://www.eol.org/>>.
- Europa Inventa*. Perth: ARC Network for Early European Research, 2010.
<<http://europa.arts.uwa.edu.au/>>.
- Europa Sacra*. Turnhout: Brepols Publishers NV, 2004–2010. <<http://www.brepolis.net/>>.
- Europeana*. Den Haag: Koninklijke Bibliotheek, 2009–2010. <<http://europeana.eu/portal>>.
- Fasti Ecclesiae Anglicanae 1066–1300*. London: University of London, Institute of Historical Research, 2003–2010. <<http://www.british-history.ac.uk/subject.aspx?subject=2&gid=39>>.
- Good, Benjamin M. and Mark D. Wilkinson. "The Life Sciences Semantic Web is full of creeps!" *Briefings in Bioinformatics* 7.3 (2006): 275–286.
- Hertel, Alice, Jeen Broekstra and Heiner Stuckenschmidt. "RDF Storage and Retrieval Systems." *Handbook on Ontologies*, ed. S. Staab and R. Studer. Berlin: Springer-Verlag, 2009 (International Handbooks on Information Systems): 489–508.
- Hitzler, Pascal, Markus Krötzsch and Sebastian Rudolph. *Foundations of Semantic Web Technologies*. Hoboken: Chapman & Hall/CRC, 2009.
- Hyvönen, Eero. et al. "CultureSampo – A national publication system of cultural heritage on the Semantic Web 2.0." *The Semantic Web: Research and Applications: ESWC 2009 Heraklion*,

- Crete, Greece, May 31–June 4, 2009 Proceedings*. Ed. Lora Aroyo et al. Berlin: Springer-Verlag, 2009 (Lecture Notes in Computer Science 5554): 851–856.
<http://www.seco.tkk.fi/publications/2009/hyvonon-et-al-culsa-demo-eswc-2009.pdf>.
- In Principio: Incipit Index of Latin Texts*. Turnhout: Brepols Publishers NV, 2004–2010.
<http://www.brepolis.net/>.
- International Medieval Bibliography*. Turnhout: Brepols Publishers NV, 2001–2010.
<http://www.brepolis.net/>.
- Isaac, Antoine and Thierry Bouchet. “Rameau et SKOS.” *Arabesques*, 54 (Apr.–June 2009), 13–14.
- Isaac, Antoine and Ed Summers, eds. *SKOS Simple Knowledge Organization System Primer*. W3C Group Note, August 18, 2009. <http://www.w3.org/TR/skos-primer>.
- Isaac, Antoine et al. “Evaluating thesaurus alignments for semantic interoperability in the library domain.” *IEEE Intelligent Systems*, 24.2 (2009), 76–86.
- Labyrinth: Resources for Medieval Studies – Manuscripts*. Washington D.C.: Georgetown University 1994–2007.
<http://labyrinth.georgetown.edu/display.cfm?Action=View&Category=Manuscripts>.
- Library of Congress Subject Headings*. Washington D.C.: Library of Congress 2009.
<http://id.loc.gov/authorities>.
- Manuscriptorium*. Praha: Národní knihovna České republiky, 2006–2010.
<http://beta.manuscriptorium.com/>.
- Medieval Manuscripts in Dutch Collections*. Den Haag: Koninklijke Bibliotheek, 2007.
<http://www.mmdc.nl/static/site/index.html>.
- MEDIUM*. Paris: Institut de recherche et d’histoire des textes, 2008.
http://www.irht.cnrs.fr/ressources/medium_frame.htm.
- Mirabile: Archivio digitale della cultura latina medievale*. Firenze: SISMEL, 2010.
<http://www.mirabileweb.it/>.
- Muzerelle, Denis. *Vocabulaire codicologique*. Paris: Institut de recherche et d’histoire des textes, 2002–2003. <http://vocabulaire.irht.cnrs.fr/vocab.htm>.
- NeuroCommons*. Cambridge, MA: Science Commons, 2009.
http://neurocommons.org/page/Main_Page.
- Pass, Gregory A. *Descriptive Cataloging of Ancient, Medieval, Renaissance, and Early Modern Manuscripts*. Chicago: Association of College and Research Libraries, 2002.
- Perseus Digital Library*. Version 4.0. Boston, MA: Tufts University, 2005–2010.
<http://www.perseus.tufts.edu/>.
- Personennamendatei* [including *Personennamen des Mittelalters*]. Frankfurt: Deutsche Nationalbibliothek, 2005–2010. <http://www.d-nb.de/eng/standardisierung/normdateien/pnd.htm>.
- RAMEAU: Répertoire d’autorité-matière encyclopédique et alphabétique unifié*. Paris: Bibliothèque nationale de France 2010. <http://rameau.bnf.fr/>.
- Schreiber, Guus et al. “Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator.” *Journal of Web Semantics*, 6.4 (2008), 243–249.
- Science Commons*. Cambridge, MA: Creative Commons, 2005–2010.
<http://sciencecommons.org/>.
- Scriptorium*. [with catalogues 1946–2008]. Bruxelles: Centre International de Codicologie 2002–2010. <http://www.scriptorium.be/>.

- Shared Names*. Boston: Science Commons, 2009–2010. <<http://sharedname.org/>>.
- Stevens, Wesley M., ed. *Bibliographic Access to Medieval and Renaissance Manuscripts: a Survey of Computerized Databases and Information Services*. Binghamton, NY: Haworth Press, 1992.
- Text Encoding Initiative. P5: Guidelines for Electronic Text Encoding and Interchange. 10 Manuscript Description*. TEI Consortium, 2007.
<<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>>.
- TextGrid*. Göttingen et al.: TextGrid Project, 2009–2010. <<http://www.textgrid.de/>>.
- Virtual Library – Historical Auxiliary Sciences – Codicology*. München: Ludwig-Maximilians-Universität, 2000–2009. <http://www.vl-ghw.lmu.de/kodikologie_en.html>.

Semantic Technologies for Manuscript Descriptions — Concepts and Visions

Robert Kummer

Abstract

The contribution at hand relates recent developments in the area of the World Wide Web to codicological research. In the last number of years, an informational extension of the internet has been discussed and extensively researched: the Semantic Web. It has already been applied in many areas, including digital information processing of cultural heritage data. The Semantic Web facilitates the organisation and linking of data across websites, according to a given semantic structure. Software can then process this structural and semantic information to extract further knowledge. In the area of codicological research, many institutions are making efforts to improve the online availability of handwritten codices. If these resources could also employ Semantic Web techniques, considerable research potential could be unleashed. However, data acquisition from less structured data sources will be problematic. In particular, data stemming from unstructured sources needs to be made accessible to Semantic Web tools through information extraction techniques. In the area of museum research, the CIDOC Conceptual Reference Model (CRM) has been widely examined and is being adopted successfully. The CRM translates well to Semantic Web research, and its concentration on contextualization of objects could support approaches in codicological research. Further concepts for the creation and management of bibliographic coherences and structured vocabularies related to the CRM will be considered in this chapter. Finally, a user scenario showing all processing steps in their context will be elaborated on.

Zusammenfassung

Der Beitrag bezieht neue Entwicklungen im World Wide Web auf die kodikologische Forschung. Seit einiger Zeit wird eine informationelle Erweiterung des Internet diskutiert und in vielen Bereichen, auch der digitalen Informationsverarbeitung des kulturellen Erbes, ausführlich erforscht und getestet: das Semantic Web. Das Konzept beinhaltet, dass Daten auf der Ebene ihrer Bedeutung miteinander verknüpft werden, damit Computer diese verarbeiten und weitere Informationen daraus gewinnen können. Im Bereich der Kodikologie gibt es schon seit einigen Jahren Bemühungen, handschriftliche Kodizes online verfügbar zu machen. Wenn auch diese den Schritt

in das Semantic Web vollziehen würden, könnten daraus nicht unerhebliche Forschungspotenziale abgeleitet werden. Die Datengewinnung aus wenig strukturierten Datenquellen ist dabei nicht unproblematisch. Insbesondere Daten aus unstrukturierten Quellen müssen zunächst mittels Verfahren der Informationsextraktion einer weiteren Verarbeitung im Sinne des Semantic Web zugänglich gemacht werden. Im Umfeld der Museumsforschung wird das CIDOC Conceptual Reference Model (CRM) ausführlich diskutiert und bereits gewinnbringend eingesetzt. Das CRM lässt sich gut auf die Forschung des Semantic Web beziehen und seine Konzentration auf Kontextualisierung von Objektzusammenhängen könnte der kodikologischen Forschung entgegenkommen. Weitere Konzepte und Standards im Umfeld des CRM zur Erstellung und Verwaltung bibliographischer Zusammenhänge und strukturierter Vokabulare werden in die Überlegungen einbezogen. Abgerundet wird die Betrachtung durch ein Benutzungsszenario, an dem verschiedene Verarbeitungsschritte in ihren Zusammenhang gestellt werden.

1. Semantic Codicology

How can the methods and tools of the Semantic Web be applied to the domain of codicology? Many handwritten codices have already been published online, mainly for viewing. Catalogs and common information retrieval techniques (e.g. full-text searching) enable discovery of information. But could additional research potential be unlocked by also making this information available according to the concepts of the Semantic Web? Could we ask and approach other questions by processing this information with the tools that have been developed in this area?

For the study and description of a specific codex, knowledge from several disciplines needs to be considered such as, for example, philology. In addition, statistical techniques have been employed to elaborate stemmata for single texts. Geographic and chronological dissemination of scripts and decorations have been considered as significant features. With regard to the individual codex, manual and technical aspects of production require study, for instance, queries regarding material (papyrus, parchment or paper), binding of folios and quires, ink and writing utensils, book decorations and provenance. Codicology simultaneously treats its research objects as material artifacts and as abstract documents. Thus, an analogy between codicology and archaeology can be drawn to a certain extent. For example, during his studies of the Rothschild collection, Delaissé showed a strong commitment to what he called “the archaeology of the book” (Delaissé, Marrow, and de Wit; Maniaci).

In order to assess the research potential of the Semantic Web for the domain of codicology, this standard should be evaluated with a focus on how methods of codicology translate into methods of Semantic Web research; explicit contextual modeling of information could be the key method that is common to both. In particular,

focusing on contextual coherences of objects, as the Semantic Web does, could support the methods of codicology.

The following passages provide an overview of Semantic Web concepts and tools. Semantic Web research itself needs to be considered as part of research in information integration and artificial intelligence. Findings in these research areas will not be exhaustively presented but rather mentioned when appropriate. Concepts and tools that have evolved as part of Semantic Web research will be introduced by an example that relates to codicology. However, no suggestions for concrete applications will be made. User scenarios have been considered helpful both for envisioning future software applications and implementing existing ideas (Alexander). An exhaustive user scenario would certainly help to understand where the ideas of the Semantic Web could support codicology in the future. Hopefully, this contribution will help to create such a user scenario for “Codicology and the Semantic Web”.

2. Semantic Web Research

Usually, information on a specific research topic is scattered among several cultural heritage information systems. In many cases, information can only be processed according to user-needs if it has been integrated. Integrated information systems can process data in a more complete fashion and usually provide better results. Additionally, they offer one consistent way of dealing with the data instead of users having to learn many user metaphors. Thus, if data stemming from different information systems is to be processed in a uniform way, it needs to be harmonized in terms of syntax and semantics. For many operations, the integrated information must reside in the main memory of a single computer to be processed efficiently and according to the needs of users.

Berners-Lee, Hendler, and Lasilla conceptualized a so-called “vision piece” that describes an infrastructure to provide greater capabilities. The authors argue that the available data on the World Wide Web has been designed for humans to read and process. They point out the importance of particular pieces of software, so-called intelligent software agents. These software artifacts are reminiscent of rational agents described by Russell and Norvig. An agent, in this sense, is designed to aid humans in information processing by acting rationally to collect, process and share data. To enable this, data currently published as part of the World Wide Web needs to be represented in certain ways. This holds true for web pages but also for databases that are considered to be part of the “deep web”.¹ Consequently, research in the area of the Semantic Web has developed from several branches of information technology: building

¹ Some web sites are generated dynamically each time a user requests a page. This part of the web is difficult to index by search engines like Google. Usually, the data is managed in some kind of proprietary structure

models, computing with knowledge and exchanging information (Hitzler, Krötzsch, and Rudolph).

In order to make information accessible for automatic processing, it has to be formalised. In the scope of the Semantic Web several concepts have been proposed. XML seems to be well established in the digitisation community, and is often used for encoding information about a codex and its contents. The Text Encoding Initiative (TEI) provides a set of XML tags for this task in chapter ten of the TEI guidelines. The Semantic Web community has proposed a recommendation that can be expressed in XML. The Resource Description Framework (RDF) can be used to express so-called triples that are simple statements which take the ordered form: ‘subject’, ‘predicate’, ‘object’. RDF is a data model that allows us to make “statements” about subjects. The World Wide Web Consortium provides an excellent introductory text on RDF (Manola and Miller). A statement like “‘De natura rerum’ is written by Beda Venerabilis” can be easily expressed as the RDF triple: “‘De natura rerum’ (subject): “is written by” (predicate): “Beda Venerabilis” (object). Subject, predicate and object may each be identified by a Universal Resource Identifier (URI). A URI is a simple string of characters that is used to identify a thing. The most commonly used form of a URI is an URL (Uniform Resource Locator), which are used daily to direct a web browser to a web page like “<http://example.org/>”. It is common practice to use URIs that have the form of URLs in the Semantic Web community to refer to a thing.

Another interesting aspect of the Semantic Web is that its community actively researches techniques from artificial intelligence. Many Semantic Web tools make use of so-called “inference engines” to deduct new knowledge from databases. Thereby, structured and sophisticated queries can rely on a larger amount of information than originally available. Furthermore, so-called “ontologies” comprise “taxonomies” and rules that can be deployed to process data according to the intended meaning.²

XML-based data, RDF, URIs and ontologies provide the tools for manipulating data as information, and even as knowledge, but can also support information sharing. With the help of ontologies different communities can agree on the meaning of certain concepts. By coordinating definitions that different communities have developed, a shared understanding of concepts can be achieved. This allows data stemming from different information systems to be processed according to the intended and agreed meaning.

that makes it difficult to share and process outside the boundaries of the system; clearly an issue that the Semantic Web tries to deal with.

² Although it has its origin in philosophy, in computer science the term “ontology” refers to a formal representation of knowledge about a certain domain. The notion of an ontology will be elaborated on in chapter 4.

3. Extracting and Modeling Information

The previous section has introduced a suite of concepts that should help to put the main ideas of the Semantic Web into practice. The central concept to encode, share and process information is the triple. So-called “Triplestores” are computer programs that control the creation, use and maintenance of data that has the form of triples. Unlike traditional relational databases, Triplestores are purpose-built for dealing with data encoded according to RDF. However, relational databases may be used to make data persistent. The RDF data model builds on the notion of a graph.³ The process of filling these stores with data will be described in this section.

It becomes apparent that the subject and object of a triple need to be atomic units of discourse that can be identified by a URI (e.g. “<http://example.org/cod/123>”: “<http://example.org/writtenIn>”: “<http://example.org/scriptorium/321>”). We want to be able to refer to exactly one concept or thing to make a statement about it. But information sources need to deliver information of high quality and granularity to establish these triples. Therefore, in some cases, information needs to be extracted rather than mapped to each data source if it does not provide enough structure.

Many information systems that deal with cultural heritage material use information retrieval methods to provide searching capabilities. In fact, traditional information retrieval tries to deal with material that is not very well-structured, such as full-text. In contrast, information extraction aims at extracting useful structured information from, for example, a text document (Konchady). In the field of Semantic Web research it would be desirable to extract triples from semi-structured sources as well as from highly structured sources (like formal descriptions of manuscripts in a catalogue). In Section 6, a user scenario will be developed that relies on highly structured data and would not be possible with traditional information retrieval. Often, information retrieval starts with identifying named entities such as people, places and institutions in documents (Cardie), not unlike a traditional printed index.

Where do we find data in the field of codicology? Many institutions have decided to publish digital information about codices according to certain standards of description. Listing 1 shows how information about a manuscript can be encoded using TEI. It is neither completely structured nor completely unstructured. Some information is highly structured, for example, the reference to material in line 12. Other information is less structured, like the information about the history of the manuscript in line 18. Inside this element some information is structured, like the reference to the archbishop

³ A graph is a concept from mathematics that is structured as sets of ordered pairs. Each pair consists of two edges that are connected by arcs. This is reminiscent of a triple where subject and object are the edges that are connected by the predicate (arc). If one edge connects to many other edges, a whole network of knowledge can emerge from simple triple statements.

of Cologne, but other information lacks precision, like the reference to a Cistercian convent.

```

1  <teiHeader>
2    <fileDesc>
3      <titleStmt>
4        <title>Biblia Sacra</title>
5      </titleStmt>
6      <sourceDesc>
7        <msDesc xml:id="kn28-0001" xml:lang="de">
8          <physDesc>
9            <objectDesc form="codex">
10              <supportDesc material="perg">
11                <support>
12                  <material>Parchment</material>
13                </support>
14              </supportDesc>
15            </objectDesc>
16          </physDesc>
17          <history>
18            <origin>Liber sancti petri a pio patre herimanno datus (<date>9/10</date>, <
              locus> f. 1r</locus> — <persName>Herimann Abp. of Koeln <date>890–923</
              date></persName>);</origin> <provenance><persName> Rutgheri </persName>. (
              <date>9– or 10c?</date>, <locus>f. 1r</locus>); <quote>Hic liber est
              sancti petri in colonia concessus conventui de prato sancte marie per
              manum domini alberti subdecani, quem idem conventus reddet sine
              contradictione, cum <sic>replitus</sic> fuerit a capitulo sancti petri,
              sicut continetur in litteris, quibus se predictus sanctimonialium
              conventus obligavit. Et in eo sunt multa folia truncata. <date>Anno MCCXII
              </date>.</quote> (<locus>f. 1r</locus>, notice dated <date>1241 </date>,
              that this book was lent by cathedral to the convent of the Prata S. Mariae
              , also called Benden; this Cistercian convent for women was founded <date>
              1207 </date> in the area of Bruehl [about 10 km south of Cologne]; [...<
              /provenance>
19          </history>
20        </msDesc>
21      </sourceDesc>
22    </fileDesc>
23  </teiHeader>

```

Listing 1. Manuscript description as part of a TEI encoded document.

What can we do about semi-structured text? Highly structured data usually can be extracted very well. The tag “<material>” indicates that the contained value will denote a certain material. A triple like “kn28-0001”, “consists of”, “Pergament” can easily be constructed if the meaning of the attribute “xml:id” is known in this context. However, the tag “<history>[...] this book was lent by cathedral to the convent of the Prata S. Mariae, [...]</history>” will be harder to extract unless a well maintained list of cathedrals and convents supports the information extraction tool.

The result of the extraction process should be a triple like “kn28-0001”, “was lent to”, “Prata S. Mariae”.⁴

Structured queries that rely on the semantics of information are only possible if the data model is also highly structured. Therefore, the migration of information to the Semantic Web cannot be limited to adopting its concepts but needs to aim at making information explicit that was implicit before. Listing 2 shows how some of the TEI information has been encoded according to Semantic Web concepts. In this case the information on the manuscript has been saved as a text file encoded in Turtle.⁵

```

1  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
2  @prefix crm: <http://erlangen-crm.org/100302/>.
3  @base <http://ceec.uni-koeln.de/>.
4
5  :kn28-0001
6      rdf:type crm:E22_Man-Made_Object;
7      crm:P1_is_identified_by [
8          rdf:value "Koln, Dombibliothek, Codex 1."@de
9      ].
10     crm:P128_carries :kn28-0001doc;
11     crm:P45_consists_of :parchment.
12
13 :kn28-0001doc
14     rdf:type crm:31_Document;
15     crm:P1_is_identified_by [
16         rdf:value "Biblia Sacra"@de
17     ];
18     crm:P1_is_identified_by [
19         rdf:value "Vulgata Bible"@en
20     ].
21
22 :parchment
23     rdf:type crm:E57_Material;
24     crm:P1_is_identified_by [
25         rdf:value "Pergament"@de
26     ];
27     crm:P1_is_identified_by [
28         rdf:value "parchment"@en
29     ].

```

Listing 2. Manuscript information modelled in Turtle.

⁴ This triple is a shortcut of a more complex set of triples that include the actor who surrendered the custody and the actor that the custody was surrendered to. A special interest group has been formed to reasearch the relation of markup like TEI to ontologies (Eide and Ore).

⁵ Turtle (Terse RDF Triple Language) is a serialization format for RDF. In this context, serialization means to dump triple data to a file for persistence or transmission. Turtle is a popular systax for RDF because it is more human-readable than XML. However, according to the recommendation, RDF should be serialized as RDF/XML (the XML syntax for expressing RDF).

Lines 1 to 3 define different namespaces that can be reused throughout the document to guarantee that keywords are unique although they have been collected from different information sources.⁶ The rest of the document can be read as an aggregation of simple subject, predicate and object statements. The information in the figure already adheres to the CIDOC CRM that will be described in the following section. Lines 5 and 6 express that there is an entity “:kn28-0001” which is an instance of the class “E22_Man-Made_Object”. Line 11 adds the information that “:kn28-0001” consists of parchment. The expression “:kn28-0001”, of course, denotes the physical codex that is part of the collection of the “Diözesan- und Dombibliothek Köln.”⁷ This information is highly structured and additional semantic information has been made explicit, thus satisfying the precondition for complex query processing.

4. The Role of Ontologies

In the field of Semantic Web research the Resource Description Framework (RDF) has been proposed as an approach to conceptually model data of a certain domain. An example of how data can be encoded according to RDF has been presented in listing 2. However, RDF does not make any recommendations as to how a certain domain could be structured or which terminology should be used. Like a traditional relational database it does not make any statements about the meaning of data, and many of the semantics have to be modeled as part of the application logic of a computer program. Ontologies have been proposed as a much richer approach to model the semantics of information. The CIDOC CRM mentioned before has been explicitly modeled as an ontology and its inventors introduced it as an “ontological approach” (Dörr).

Information technology took the word “ontology” from philosophy in an analogy but redefined the term to fit its needs. In fact, ontologies have been considered as being a “silver-bullet” for information integration (Fensel). Basically, ontologies have been introduced to support communication processes in larger groups. They have been developed to help organisations find a common language and understanding of important domain concepts. In comparison to flat glossaries or terminology lists, ontologies comprise a complex thesaurus-like structure, additional rules and

⁶ Because it would be cumbersome to write the full URI of each part of the triple, so-called namespaces can be defined. A namespace definition binds a part of the full URI to a qualified name that can be used throughout the document. The base namespace is applied to all names that omit the qualified name in front of the colon connecting the qualified name with its suffix. For example, “:kn28-0001” translates to “<<http://ceec.uni-koeln.de/kn28-0001>>” and “rdf:value” to “<<http://www.w3.org/1999/02/22-rdf-syntax-ns#value>>”.

⁷ The digital facsimiles of the “Diözesan- und Dombibliothek Köln” have been published as “Codices Electronici Ecclesiae Coloniensis” (Thaller and Finger). “kn28” denotes the identification code of the “Diözesan- und Dombibliothek”.

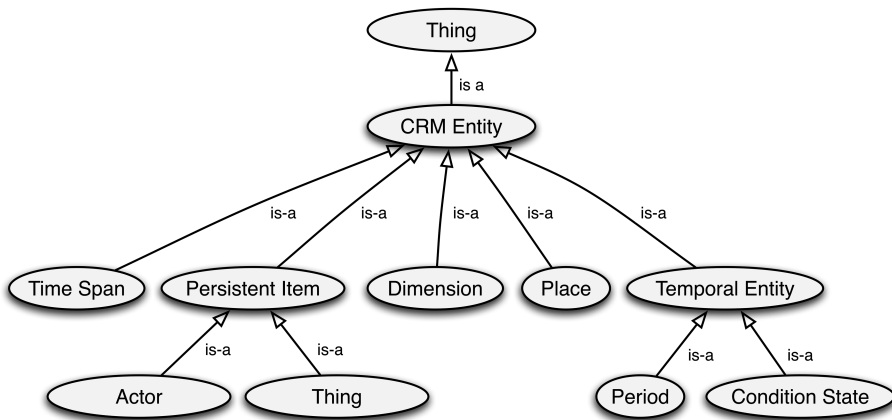


Figure 1. The Class hierarchy of the CIDOC CRM.

restrictions.⁸ Thus, they not only define certain notions but can also encode complex interrelations. Although there is no canonical definition of what an ontology is in information technology, Gruber was the first to formalise the topic. Ontologies can be constructed with the Web Ontology Language (OWL).⁹

In 2006, the CIDOC Conceptual Reference Model was accepted as official standard ISO 21127:2006 (Crofts et al.). It provides a taxonomy for expressing information about material objects in the cultural-heritage area. Like any ontology, it can be used both to support communication processes in larger communities that strive for sharing of information and to implement software systems that integrate information from different information systems. A hierarchy of classes defines concepts that are commonly referred to in museum documentation practice. And so-called properties form relations between these conceptual classes. Up to now, the CRM has been used in several integration projects.¹⁰

Figure 1 shows a part of the class hierarchy provided by the CIDOC CRM. The visualization has been generated from an OWL implementation of the “Erlangen CRM”

⁸ For example, a rule that states the uncle relationship in a fictional family ontology can have the form [rule1: (?f pre:father ?a) (?u pre:brother ?f) -> (?u pre:uncle ?a)] (the rule is written in the syntax of the Jena Semantic Web framework; the example is taken from <<http://jena.sourceforge.net/inference/#rules>>). Rules are evaluated and processed by inference engines to create new facts (triples). An example of a restriction is that a human always has, at most, two arms.

⁹ More information about OWL can be found at Smith, Welty, and McGuinness.

¹⁰ Two examples would be SCULPTEUR (Giorgini) and BRICKS.

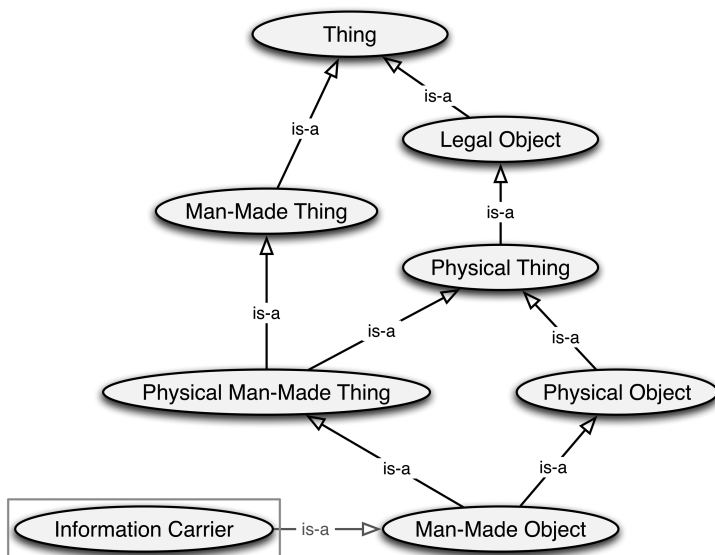


Figure 2. An information carrier in the CRM hierarchy.

(Schiemann et al.). For modeling data in the field of codicology, the CRM provides both classes for modeling a codex (for example, as a physical man-made object or information carrier) and the contained work (for example, as a document). But also classes for describing additional contextual information such as the condition of a codex, people involved in its creation and the history of ownership. The following section will introduce two relevant classes: “E84_Information_Carrier” and “E31_Document”.

In order to describe a codex as a material thing, for example, one might use the CRM class “E84.Information Carrier”. From the official CIDOC CRM documentation: “This class comprises all instances of E22 Man-Made Object that are explicitly designed to act as persistent physical carriers for instances of E73 Information Object. This allows a relationship to be asserted between an E19 Physical Object and its immaterial information contents” (Crofts et al. 67). Figure 2 shows the class as part of the inheritance hierarchy of the CRM. It is important to keep in mind that each class inherits all the features of its super class.

The contained textual material considered as a conceptual object can be modeled as “E31.Document”. The official documentation defines that this class “comprises identifiable immaterial items, which make propositions about reality. These propositions may be expressed in text, graphics, images, audiograms, videograms or by other similar means”

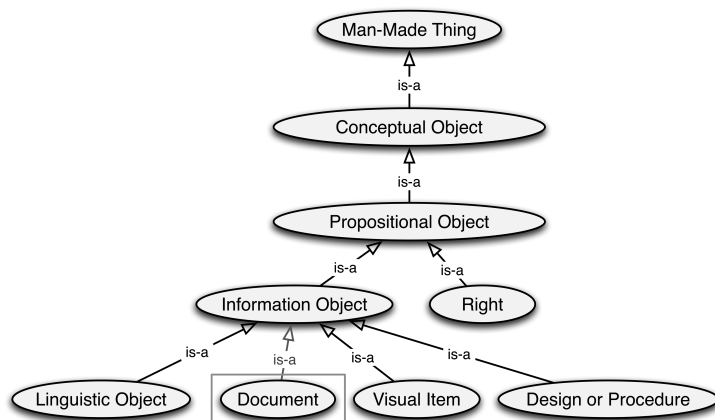


Figure 3. A document in the CRM hierarchy.

(Crofts et al. 48). Figure 3 shows how a document according to the CRM is also a man-made thing but without material character. However, it would be beyond the scope of this contribution to discuss if a document in this sense can describe the features of a certain hand writing or whether another class like “E36.Visual_Item” would be a better fit for this.

The structure of the CIDOC CRM relies heavily on the notion of events. Dörr and Kritsotaki argue that modeling events in metadata is helpful for dealing with cultural heritage information. For example, the notion of an event can be helpful for expressing uncertain information. The class “E13.Attribute_Assignment”, which is a sub-class of “E7.Activity”, has been provided to emphasize how a statement about something came about. Opinions of different authors can be distinguished by using “E13.Attribute_Assignment” for each researcher’s assertion about a codex. Additionally, the history of ownership of a codex can be modeled by using events. Although developed in a museum context, classes like “E10.Transfer_of_Custody” and “E8.Acquisition” (also both sub-classes of “E7.Activity”) suggest that the CRM provides structures that can be adapted to the needs of research in the field of codicology.

But how do codices relate to the contained works in the world of CIDOC CRM? The class hierarchies shown in figures 1, 2 and 3 do not display the properties mentioned above which are needed in order to relate instances of these classes to each other. Figure 4 highlights another perspective. Instead of the class hierarchy, the relations between

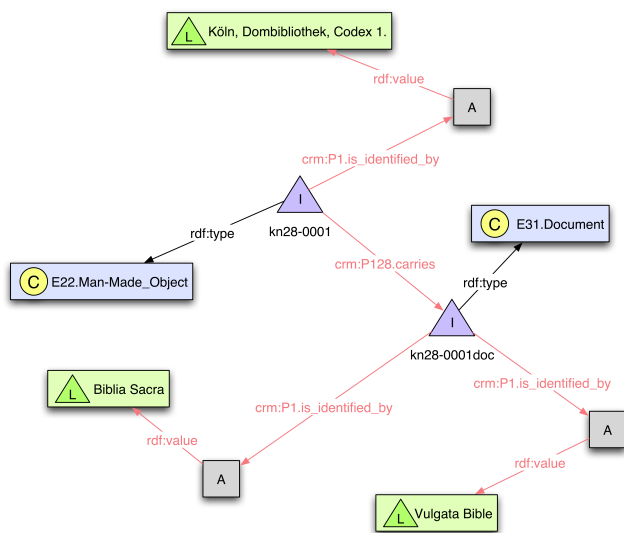


Figure 4. Manuscript information graph visualization.

the individuals are presented. Please note that this visualization has been automatically generated from the Turtle code in listing 2.¹¹

Two further developments worth mentioning that deal with structured vocabularies and bibliographies are SKOS and FRBRoo. Up to now, the discussion has focused on integrating different data models and schemas. However, different groups tend to refer to the same thing by different names. Just think of an international research environment where codex materials are referred to using national languages (e.g. “Papier”, “paper” and “páipear”). Lines 22 to 29 in listing 2 demonstrates how two different names have been assigned to the URI “:parchment”. Here the URI denotes the material itself and the different names have been associated by using the CRM property “`crm:P1_is_identified_by`”. One way to approach the terminology problem is to provide structured controlled vocabularies by using SKOS (the Simple Knowledge Organization System). It is a family of formal languages designed for any type of structured controlled vocabulary (Miles and Bechhofer). While CIDOC CRM is a formalisation of how cultural heritage content can be encoded, SKOS is a formalisation of how structured terminologies can be encoded. Figure 5 illustrates how appellations of different materials can be expressed according to SKOS. It shows that the material which

¹¹ RDF Gravity has been used to generate the visualization (Goyal and Westenthaler). For better readability the figure has been reworked by using a charting tool.

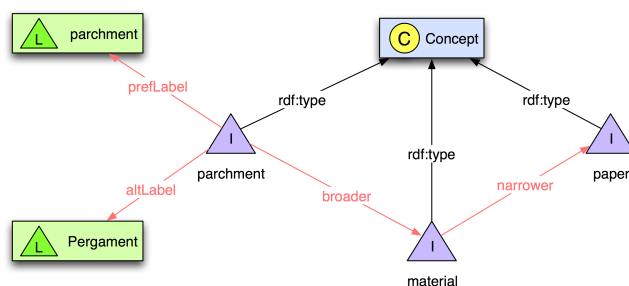


Figure 5. A graphical visualization of a vocabulary as SKOS.

the URI “:parchment” refers to has been associated with the SKOS class “Concept”.¹² The hierarchical links “:broader” and “:narrower” indicate that “:material” is more general than “:parchment” and “:paper”. Of course, SKOS data is not valuable in itself but needs to be made available for information systems so that they can make use of the structured vocabularies.

Another standard with a strong connection to Semantic Web research has been proposed in the field of library science. The Functional Requirements for Bibliographic Records (FRBR) form a conceptual model developed by the International Federation of Library Associations Institutions (Tillett). FRBR distinguishes the notions *Work*, *Expression*, *Manifestation* and *Item* (IFLA Study Group on FRBR). According to the definition document, a *Work* is an intellectual creation (for example “Moby Dick”) and the *Expression* is the realization of this creation in its distinct form (e.g. German translation of “Moby Dick”). A *Manifestation* is “the physical embodiment of an expression of a work. As an entity, manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form” (e.g. a certain edition of the German translation). Finally, an *Item* is “a single exemplar of a manifestation. The entity defined as item is a concrete entity” (a certain copy of a certain edition). A medieval codex would be defined as an *Item* in the terminology of FRBR. And since for each manifestation there is just one item, the distinction between *Manifestation* and *Item* does not seem to be pertinent to practical research in the field of codicology. Although FRBR is powerful at modeling relations between these four layers, it does not come with the means to express the history of development for an old manuscript.

¹² The SKOS reference document defines the class “Concept”: “A SKOS concept can be viewed as an idea or notion; a unit of thought. However, what constitutes a unit of thought is subjective, and this definition is meant to be suggestive, rather than restrictive” (Bechhofer and Miles).

FRBR has been harmonised with the CIDOC CRM. Therefore, it has been expressed as a formal ontology that links to the classes of the CRM. The harmonisation project strives for better related representations of bibliographic and museum information, and to facilitate the “integration, mediation and interchange” of information (Dörr and Le Boef). In June 2009 the latest version of the standard was released (Aalberg et al.).

5. Putting it Together

We have looked at how different cultural heritage information systems can use information extraction (for unstructured and semi-structured texts) and mapping (for databases or the “deep web”) to get a grip on relevant entities. We have discussed how these entities and their relations can be modeled as RDF triples in a way that conforms to a standard like the CIDOC CRM. Adhering to this standard enables software to process data according to the intended meaning. It is helpful for further data fusion tasks and complex querying of information to integrate the information acquired from different sources in one physical place. This enables comprehensive mining, indexing and querying.

Different architectures have been developed to tie together complex distributed systems ranging from distributed databases over middleware software to Service-Oriented Architectures with Web Services.¹³ In the cultural heritage domain it has become very common to publish information using the OAI Protocol for Metadata Harvesting (OAI-PMH), which relies on the HTTP protocol. The OAI-PMH has been suggested by the Open Archives Initiative for publishing and collecting metadata. Data providers, such as archives, publish their metadata as XML and service providers harvest that data in order to offer further services (Lagoze et al.). Using this protocol it would be possible to publish both the TEI documents and additional semantic data in RDF. Recent suggestions in Semantic Web research tend to avoid cumbersome approaches in favor of light-weight infrastructures. In 2006, Berners-Lee articulated his thoughts on a concept that he called “Linked Data”. The core of this concept is not only to put data on the web in a certain manner but also to link related data. Since then, the notion of Linked Data refers to a set of best practices for publishing and connecting structured data on the World Wide Web. While the concept of Linked Data requires service providers to systematically crawl linked data to acquire information, OAI-PMH offers guided and streamed pulling of data.

¹³ A Service-Oriented Architecture provides loosely coupled software components that provide services. These services can be combined to solve certain tasks. A Web Service provides an application programming interface that can be called via the HTTP protocol. HTTP (Hypertext Transfer Protocol) is a standard that is used to transmit data over a network (in particular for the Internet). Therefore it is ubiquitously available.

Once the data has been integrated, tools are needed for further processing. Semantic Web Frameworks like Jena (Carroll et al.) support software developers in creating Semantic Web applications. They provide an integrated set of tools that facilitate the design and operation of a knowledge base that can deal with triples. Most Semantic Web frameworks provide a so-called SPARQL endpoint to query the knowledge base. SPARQL is an acronym that stands for “SPARQL Protocol and RDF Query Language” (Prud’hommeaux and Seaborne). An endpoint is an entry point for a service that can be called over a network. SPARQL allows the formulation of queries that are highly structured. It does have some similarities to SQL.¹⁴ But while SQL is commonly used to query or manipulate data in a relational model, SPARQL can be used to formulate complex graph-like query structures to query RDF data.¹⁵ SPARQL queries can be transmitted over the HTTP protocol. Larger projects are picking up the idea of the Semantic Web and developing advanced applications. Information extraction projects like DBpedia mine the World Wide Web for structured data and make it accessible for the Semantic Web (Auer et al.). Another example is the SIMILE project (Mazzocchi, Garland, and Lee). It is more end-user-oriented than DBpedia and develops tools that examine the possibility of semantically processing digital assets.

6. User scenario

The Semantic Web has been introduced and codicological material referenced, only the coherent user scenario remains to be described. Therefore, a simple use case for further elaboration will be presented here. It will develop around a simple question and highlight the implications of how Semantic Web technology will be affected. In the above-mentioned vision piece a user scenario has been developed that should motivate future research and funding in the area, but scenarios can also be used in the microcosm of system development. They help to reflect on functional requirements that a single piece of software needs to fulfill in order to meet a user’s needs. User scenarios facilitate communication between software designers, programmers and end-users by providing a shared example.

Imagine a researcher working on medieval codices. To push forward his current research project, he is interested in how texts spread in certain institutions in a specific region.¹⁶ He has a good friend in the IT department of the university who enthusiastically reported on a new strain of research, the Semantic Web. According to this concept, digital information will be managed in a way that supports machine-

¹⁴ SQL is ISO standard ISO/IEC 9075:2008 and stands for Structured Query Language.

¹⁵ Up to now there is no W3C recommendation for the manipulation capabilities of SPARQL. However, most Semantic Web toolkits provide capabilities for data manipulation outside SPARQL and extensions for SPARQL are being developed.

¹⁶ I want to thank Almut Breitenbach and Patrick Sahle for their support in creating this user scenario.

processing. The researcher wonders if this new technology could meet a requirement he has formulated as follows: “For the geographic area of northern Germany, show all codices that contain texts of Classic Latin authors and that have been written in the 13th century. Draw the results as circles on a map and use different colors for monasteries and nunneries.” Many requirements need to be fulfilled to enable a system to process such a question.

Certainly, the data that is needed to compile the results resides on scattered information systems, preferably encoded as structured manuscript descriptions. As a first step, this data needs to flow from one information system to another. Catalogue data from different information systems has been published and is exposed via OAI-PMH as TEI. Imagine an information system that strives to support the researcher. It will request information from several data providers and gather it for further processing. This approach requires little effort for data providers. Other architectures could demand that one or more of the following pre-processing steps be performed by data providers before the data is published. As a first step, information extraction needs to be performed on the acquired data by the central information system. The system aims to extract named entities and to assign the right unique identifier (i.e. URI) to each entity. This step is rather important because without canonical names across the participating information systems all following steps will fail.

Once entities and the relations that exist between them are represented as URIs, they can be stored as triples in some serialisation of RDF. To be available for processing, triples are held in main memory according to a suitable data structure. For exchanging information between different information systems, this data needs to be serialised in a file. An example for a serialisation has been given in listing 2. After ingesting the triples in a triplestore, the data will be available for further processing.

The extracted entities alone are of very limited use unless they are aligned with additional background knowledge. This knowledge will be provided by specialised knowledge bases as triples. It comprises, for example, the geographic region, the monastery and religious order mentioned in the manuscript description. Without this background, none of which is contained in the metadata of the codex alone, the query of the researcher cannot be answered. But after adding the supplementary knowledge to the triplestore, additional facts are available that can be considered for query processing. For example, the three triples “codex123”, “carries”, “document123”, “document 123”, “has author”, “Cicero” (both extracted from codex information) and “Cicero”, “has genre”, “Classical Latin work” (added from background knowledge base) can be combined to reason that the codex contains a text of an author that has been attributed “Classical Latin work”. Additional facts can be derived by applying rules. And plausibility checks can be conducted to disclose contradictory information that may emerge by considering additional knowledge.

We assume that the information about the author of a text and its place of creation could be extracted from the codex metadata. Additionally, a group of theoretical researchers recorded their findings by putting results in a specialized information system (for example that texts of a certain author usually can be ascribed to a specific genre). Another system contributes the geographic coordinates for a certain geographic region. If the information system that the researcher is using has access to all the above systems, they can now formulate queries that could not have been formulated before. Listing 3 shows a selected aspect of the aforementioned query in a formalised way. Its formalisation is little more than preliminary but seems to be sufficient to discuss the process of formalisation.

```

1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX crm: <http://erlangen-crm.org/100302/>
3  PREFIX cod: <http://codicology.org/>
4  PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
5  BASE <http://ceec.uni-koeln.de/>
6
7  SELECT ?codex, ?gender, ?geo
8  WHERE {
9    ?codex rdf:type crm:E22_Man-Made_Object .
10   ?codex crm:P108I_was_produced_by ?codexProduction .
11   ?codexProduction crm:P7_took_place_at ?monastery .
12   ?monastery crm:P2_has_type ?gender .
13   ?gender skos:broader cod:gender .
14   [...]
15 }
```

Listing 3. A SPARQL query that formalizes a research problem.

As in listing 2, the query starts by defining a couple of namespaces including the base namespace to ensure that the used names are unique. The rest of the query is based on the syntax of SPARQL that serves to query triplestores. Then, the three variables “?codex”, “?gender” and “?geo” are defined to carry the results. The result will hold identifiers of codices together with the information about the geographic unit the monastery belongs to and whether it is a nunnery or monastery. The following part of the query re-uses these variables and defines additional ones that are only used temporarily. The temporary variables like “?monastery” are needed to build the “factual bridges” from one piece of information to another. The property “crm:P2.has_type” connects the “?monastery” with its “?gender”. Multiple types can be connected with a “crm:E22.Man-Made_Object” and therefore the property “skos:broader” has been used to restrict the value of the assigned type to be a specialisation of “cod:gender”, which could be either male or female.

For real world use a graphical user interface needs to be implemented. The query demonstrated in listing 3 could not be formulated by a researcher in the field of codicology or palaeography without undertaking significant additional training. On

the foundation of the mentioned Semantic Web framework, additional software layers need to mediate between the end-user and the bits and bytes. Many questions arise when thinking about such a system like questions of helpful interface design for easy interaction and formulation of very complex queries. Research in artificial intelligence strives for processing queries in natural language. The Companions Project, for example, explores software that is reminiscent of the intelligent agents mentioned before (Benyon and Mival).¹⁷

The information that has been acquired by evaluating the example query has the form of a table with the columns holding an identifier for the codex, the information if it is a monastery or nunnery and its geographic region. Although this information is helpful for the researcher it would be useful to display the results as a map. With the advent of Web 2.0, mashups have become quite popular.¹⁸ The fictional researcher could use a similar service to display a map of the region of his interest. The web application would draw a circle for each monastery on the map with different colors for monasteries and nunneries. By querying a geographic database, the geographic identifier can be resolved to coordinates that are needed for the drawing task.

It is obvious that only a few of the infrastructural elements which would facilitate the user scenario are available today. Scientists in the field of codicology would need to make their factual knowledge available as a domain-specific knowledge base. Additionally, dealing with consistent and canonical URIs is anything but easy. Information extraction usually can only extract entities that it can look up in some kind of authority file or that it has been trained to find by machine learning techniques. Other entities can be identified but not resolved to a canonical name, especially in the case of unstructured text. Another problem is the scalability of current Semantic Web Triplestores. Unlike relational databases they are still not well understood and cannot deal with massive amounts of data. However, the Semantic Web community has recognised this problem and is working on scalable solutions. One example is the OWLIM Semantic Repository (Ontotext AD) that scales to several billion triples. The W3C maintains a Wiki that lists Triplestores, sorted by their scalability.

¹⁷ The aforementioned SIMILE project is experimenting with different user interfaces that provide faceted browsing, timelines and maps. Another example would be PhiloSpace, a piece of software that has been developed within the COST framework. It can be used to establish semantic relations among entities of the philosophical domain.

¹⁸ Among other things, the concept of Web 2.0 means that users can interact with the web site and actively contribute to it. Mashups are one example for a web page that combines data and functionality from other resources to create a custom service.

7. Concluding Remarks

This contribution aims to introduce key concepts and tools of Semantic Web research. It does not claim to articulate future directions of research but tries to provide criteria and background information for researchers which may help with their decision processes. Semantic Web technology certainly offers new perspectives on how data will be published, shared and processed in the future. However, the concepts of Semantic Web research have also been criticised. The idea that was formulated in 2001 has not yet been realised (Shadbolt, Berners-Lee, and Hall). Consequently, doubts remain about the practical feasibility of the concept because it is resource consuming to create knowledge bases and to add comprehensive structure to data. With billions of facts that have been published as triples on-line there will be scalability issues. Projects like the Large Knowledge Collider are exploring reasoning with incomplete knowledge due to limited resources (Fensel). Large ontologies tend to be cumbersome and difficult to understand, RDF with its explicit mode of expression becomes verbose and bulky. It has also been argued that the Semantic Web is not semantic. Gärdenfors for example doubts the practical feasibility of the Semantic Web because of its focus on formal syllogisms that stem from formal logic and research in artificial intelligence. These cover only a (dispensable) fraction of semantic operations that scientists want to have performed on their data. And although URIs are proposed as a unique way of identifying things, no data provider is forced to use canonical URIs. The database community can look back on a long research tradition in information integration that could (and already does) contribute valuable input (Leser and Naumann).

However, the vision of the Semantic Web has promoted a plethora of research projects in different domains (some of them mentioned in this contribution). Because of the data model that can represent data with rich and varying structure, it seems to be well suited for the humanities. Since RDF relies on the notion of a graph as its data model, it facilitates the construction of semantic networks of huge complexity and high flexibility. Cultural heritage information models often “suffer” from relying on inflexible structures that do not explicitly model the intended meaning of information objects. Again, this could limit the opportunities for helpful applications. Additionally, RDF handles missing data very well, the concept relies on the “open world assumption”.¹⁹

Thus, the Semantic Web seems to be both a blessing and a curse for information integration and processing in cultural heritage. It envisions new and interesting approaches that could be very useful for humanities information science. But research projects cannot just adopt the concepts and hope for the best. These projects should be

¹⁹ The “open world assumption” is used in knowledge representation because nobody can comprehensively model the knowledge of a certain domain. By that, one has to conclude that a software system needs to deal with incomplete knowledge and that the kinds of inference which a piece of software can perform are limited to those statements which are available.

prepared to actively engage in Semantic Web research and adequate resources should be allocated (fortunately, a very lively field at the moment). Its data model seems to be well suited to encode codicological data but its mechanisms for manipulating that knowledge seem to be restricted to formal syllogisms. Provided that the means to deal with uncertain and contradictory information are developed, the Semantic Web could foster research in areas that heavily rely on qualitative data, such as codicology. So far, many projects in the field of “humanities information science” focus on encoding information to make it available to a greater audience for searching and browsing. Manipulation of data as the primary method to generate significant insights seems to be restricted to problems that are clearly quantifiable or that can be dealt with by statistical analysis. Certainly, the Semantic Web will not provide for all the information needs of a researcher, but it could begin to play out its strength in well defined and carefully bounded applications.

Bibliography

- Aalberg, Trond et al. *FRBR - Object-Oriented Definition and Mapping to FRBRer*. International Working Group on FRBR and CIDOC CRM Harmonisation, 1.0 ed., 2009.
- Alexander, Ian and Neil Maiden. *Scenarios, Stories, Use Cases: Through the Systems Development Life-Cycle*. John Wiley & Sons, 2004.
- Auer, Sören. et al. “DBpedia: A Nucleus for a Web of Open Data.” *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*. eds. Karl Aberer et al., vol. 4825, chap. 52. Berlin, Heidelberg: Springer, 2007. 722–735.
- Benyon, David and Oli Mival. “Introducing the Companions Project: Intelligent, Persistent, Personalised Interfaces to the internet.” *BCS-HCI '07: Proceedings of the 21st British HCI Group Annual Conference on HCI 2008*. Swinton, UK: British Computer Society, 2007, 193–194.
- Berners-Lee, Tim. “Linked Data - Design Issues.” World Wide Web Consortium 2006-2009. <<http://www.w3.org/DesignIssues/LinkedData.html>>.
- Berners-Lee, Tim, James Hendler, and Ora Lassila. “The Semantic Web.” *Scientific American* 284 (2001).5: 34–43.
- BRICKS. “BRICKS Project. Building resources for Integrated Cultural Knowledge Services.” Bricks Project 2004-2007. <<http://www.brickcommunity.org>>.
- Cardie, Claire. “Empirical Methods in Information Extraction.” *AI Magazine* 18 (1997).4: 65–80.
- Carroll, Jeremy J. et al. “Jena: Implementing the Semantic Web Recommendations.” *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. New York, NY, USA: ACM, 2004, 74–83.
- COST. “Working Group 2: Software”. COST Action 32 2009. <<http://www.cost-a32.eu/wg-2.html>>.
- Crofts, Nick. et al. “Definition of the CIDOC Conceptual Reference Model.” ICOM/CIDOC 2003-2005. <http://www.cidoc-crm.org/docs/cidoc_crm_version_4.2.pdf>.

- Delaissé, Leon M. J., James H. Marrow, and John de Wit. "Illuminated Manuscripts: The James A. de Rothschild Collection at Waddesdon Manor." Fribourg: Office du Livre [et al.], 1977.
- Dörr, Martin and Athina Kritsotaki. "Documenting Events in Metadata." *The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2006)*. eds. M. Ioannides et al. 2008.
- Dörr, Martin. "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata." *AI Mag* 24 (2003).3: 75–92.
- Dörr, Martin and Patrick Le Boeuf. "FRBRoo Introduction." ICOM/CIDOC 2009.
<http://www.cidoc-crm.org/frbr_inro.html>.
- Eide, Øyvind and Christian-Emil Ore. "SIG:Ontologies." Text Encoding Initiative 2004-2010.
<<http://wiki.tei-c.org/index.php/SIG:Ontologies>>.
- Fensel, Dieter. "Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce." Springer-Verlag New York, Inc., 2003.
- Fensel, Dieter et al. "Towards LarKC: A Platform for Web-Scale Reasoning." *ICSC*. 2008, 524–529.
- Gärdenfors, Peter. "How to Make the Semantic Web More Semantic." *Formal Ontology in Information Systems: proceedings of the third international conference (FOIS-2004)*. eds. Achille C. Varzi and Laure Vieu, vol. 114 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2004, 17–34.
- Giorgini, Fabrizio. "SCULPTEUR (Semantic and content-based multimedia exploitation for European benefit)." *Sculpteur* 2002-2005. <<http://www.sculpteurweb.org>>.
- Goyal, Sunil and Rupert Westenthaler. "RDF Gravity (RDF Graph Visualization Tool)." Salzburg: Salzburg Research Forschungsgesellschaft 2004.
<<http://semweb.salzburgresearch.at/apps/rdf-gravity>>.
- Gruber, Thomas R. "A Translation Approach to Portable Ontology Specifications." *Knowl. Acquis.* 5 (1993).2: 199–220.
- Hitzler, Pascal, Markus Krötzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies*. London: Chapman & Hall/CRC, 2009.
- IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records - Final Report*. UBCIM Publications - New Series Vol 19. K. G. Saur München, 2008.
- Konchady, Manu. *Text Mining Application Programming*. Boston, Mass.: Charles River Media, 2006.
- Lagoze, Carl et al. "Open Archives Initiative Protocol for Metadata Harvesting." Open Archives Initiative 2002-2008. <<http://www.openarchives.org/OAI/openarchivesprotocol.html>>.
- Leser, Ulf and Felix Naumann. *Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen*. Heidelberg: dpunkt-Verl., 2007, 1. ed.
- Maniaci, Marilena. *Archeologia del manoscritto: metodi, problemi, bibliografia recente*. I libri di Viella. Roma: Viella, 2002, 1. ed.
- Manola, Frank and Eric Miller. "RDF Primer." World Wide Web Consortium 2004.
<<http://www.w3.org/TR/rdf-primer>>.
- Mazzocchi, Stefano, Stephen Garland, and Ryan Lee. "SIMILE: Practical Metadata for the Semantic Web." O'Reilly 2005. <<http://www.xml.com/pub/a/2005/01/26/simile.html>>.

- Miles, Alistair and Sean Bechhofer. "SKOS Simple Knowledge Organization System Reference." World Wide Web Consortium 2009. <<http://www.w3.org/TR/skos-reference>>.
- Norvig, Peter and Stuart Russell. *Artificial Intelligence: A Modern Approach*. Prentice Hall International, 2003, 2. ed.
- Ontotext AD. "OWLIM Semantic Repository." Ontotext 2010. <<http://www.ontotext.com/owlim>>.
- Prud'hommeaux, Eric and Andy Seaborne. "SPARQL Query Language for RDF." World Wide Web Consortium 2008. <<http://www.w3.org/TR/rdf-sparql-query>>.
- Schiemann, Bernhard et al. "Short Documentation of the CIDOC CRM (4.2.4) Implementation in OWL-DL." Erlangen: Friedrich-Alexander-Universität Erlangen 2008. <http://erlangen-crm.org/docs/documentation_crm_owl-dl_4.2.4.pdf>.
- Shadbolt, Nigel, Tim Berners-Lee, and Wendy Hall "The Semantic Web Revisited." *Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications]* 21 (2006).3: 96–101.
- Smith, Michael K., Chris Welty and Deborah L. McGuinness. "OWL Web Ontology Language Guide." World Wide Web Consortium 2004. <<http://www.w3.org/TR/owl-guide>>.
- TEI Consortium. "Text Encoding Initiative." Text Encoding Initiative 2010. <<http://www.tei-c.org/index.xml>>.
- Thaller, Manfred and Heinz Finger. "Codices Electronici Ecclesiae Coloniensis (CEEC)." Köln, Universität zu Köln 2002. <<http://www.ceec.uni-koeln.de>>.
- Tillett, Barbara B. "What is FRBR?: A Conceptual Model for the Bibliographic Universe." Washington (DC): Library of Congress 2004. <<http://www.loc.gov/cds/downloads/FRBR.PDF>>.
- W3C. "LargeTripleStores - ESW Wiki." World Wide Web Consortium 2010. <<http://esw.w3.org/LargeTripleStores>>.

Handschriften und Naturwissenschaften



Manuscripts and the Sciences

Automatic Palaeographic Exploration of Genizah Manuscripts

Lior Wolf, Nachum Dershowitz, Liza Potikha, Tanya German, Roni Shweka,
Yaacov Choueka

Abstract

The Cairo Genizah is a collection of hand-written documents containing approximately 350,000 fragments of mainly Jewish texts discovered in the late 19th century. The fragments are today spread out in some 75 libraries and private collections worldwide, but there is an ongoing effort to document and catalogue all extant fragments. Palaeographic information plays a key role in the study of the Genizah collection. Script style, and—more specifically—handwriting, can be used to identify fragments that might originate from the same original work. Such matched fragments, commonly referred to as “joins”, are currently identified manually by experts, and presumably only a small fraction of existing joins have been discovered to date. In this work, we show that automatic handwriting matching functions, obtained from non-specific features using a corpus of writing samples, can perform this task quite reliably. In addition, we explore the problem of grouping various Genizah documents by script style, without being provided any prior information about the relevant styles. The automatically obtained grouping agrees, for the most part, with the palaeographic taxonomy. In cases where the method fails, it is due to apparent similarities between related scripts.

Zusammenfassung

Die Geniza aus Kairo ist eine Sammlung von ca. 350.000 handschriftlichen Fragmenten jüdischer Texte, die im späten 19. Jahrhundert entdeckt wurde. Die Fragmente werden heute in 75 Bibliotheken und privaten Sammlungen auf der ganzen Welt aufbewahrt. Eine umfassende Dokumentation und Katalogisierung aller Fragmente ist in Arbeit. Paläographische Informationen spielen für die Erforschung der Geniza eine entscheidende Rolle. Schriftstil und Handidentifizierung können genutzt werden, um Fragmente der selben Quelle zu identifizieren. Solche zusammengehörigen Fragmente (sog. “joins”) müssen zur Zeit manuell von Experten gefunden werden. Es ist davon auszugehen, dass nur eine kleine Zahl solcher “joins” bis heute entdeckt werden konnte. In diesem Beitrag sollen zuverlässige Methoden zur automatischen Identifikation von Händen vorgestellt werden, die auf unspezifischen Merkmalen beruhen und einen Corpus von Schriftbeispielen benutzen. Zusätzlich untersucht der

Beitrag Möglichkeiten, Geniza-Dokumente nach der Schriftart zu klassifizieren. Diese automatisch erschlossenen Gruppen stimmen größtenteils mit einer paläographischen Taxonomie überein. In einzelnen Fällen scheitert die Methode auf Grund offensichtlicher Ähnlichkeiten zwischen den Schriftarten.

1. Introduction

Written text is one of the best sources for understanding historical life. Community documents, religious works, personal letters, and commercial records can all contribute to a better understanding of a given place and time. In this respect, the Cairo Genizah is a unique treasure trove of middle-eastern texts, comprising some 350,000 manuscripts fragments, written mainly in the 10th to 15th centuries. Discovered in the 1890s in the attic of a synagogue in Fostat, an old quarter of Cairo, the Genizah is a large collection of discarded codices, scrolls, and documents. It contains a mix of religious Jewish documents with a smaller proportion of secular texts. With few exceptions, these documents are made of paper or parchment, and the texts are written mainly in Hebrew, Aramaic, and Judeo-Arabic (Arabic language in Hebrew characters), but also in many other languages (including Arabic, Judeo-Spanish, Coptic, Ethiopic, and even one in Chinese).

After its discovery, the Genizah attic was emptied in several stages. The bulk of the material was obtained by Solomon Schechter for Cambridge University, but there were various acquisitions by others, too. By now, the contents have found their way to over 75 libraries and collections around the world. Most of the items recovered from the Cairo Genizah have been microfilmed and catalogued in the intervening years, but the photographs are of mediocre quality and the data incomplete, with thousands of fragments still not listed in published catalogues.

Genizah documents have had an enormous impact on 20th-century scholarship in a multitude of fields, including Bible, rabbinics, liturgy, history, and philology. The major finds include fragments of lost works (such as the Hebrew original of the apocryphal Book of Ecclesiasticus), fragments of hitherto unknown works (such as the Damascus Document, later found among the Qumran scrolls), and autographs by famous personages, including the Andalusians Yehuda Halevi (1075–1141) and Maimonides (1138–1204). Genizah research has, for example, transformed our understanding of medieval Mediterranean society and commerce, as evidenced by S. D. Goiten's monumental five-volume work, *A Mediterranean Society*.¹

The philanthropically-funded Friedberg Genizah Project, headquartered in Jerusalem, is in the midst of a multi-year process of digitally photographing (in full color, at

¹ See Reif for the history of the Genizah and of Genizah research.

600dpi) most—if not all—of the extant manuscripts. The entire Genizah collections of the Jewish Theological Seminary in New York (ENA), the Alliance Israelite Universelle in Paris (AIU), The Jewish National and University Library in Jerusalem (JNUL), the recently rediscovered collection in Geneva, and many smaller collections have already been digitized and comprise about 90,000 images (recto and verso of each fragment). The digital preservation of another 140,000 fragments of the Taylor-Schechter Genizah Collection at The Cambridge University Library is currently underway. At the same time, everything that is known about the fragments is being extracted from books, catalogues, and scholarly articles. The images and all the information about them are made freely available to researchers online at www.genizah.org.

Late in 2008, the Friedberg Genizah Project embarked on an ambitious effort to apply the latest image-processing technology and artificial-intelligence research to the analysis of its archive of images, thereby providing scholars of the humanities with new and powerful tools for Genizah research. This work is being carried out in cooperation with computer-science researchers in the fields of vision and machine learning from Tel Aviv University, the Hebrew University of Jerusalem, and Ben-Gurion University of the Negev and in consultation with palaeographers and Genizah scholars. We report on some aspects of that endeavor here.

Consider that, unfortunately, most of the leaves that were found were not found in their original bound state. Worse, many are fragmentary, whether torn or otherwise mutilated. Pages and fragments from the same work (book, collection, letter, etc.) may have found their way to disparate collections around the world. Some fragments are very difficult to read, as the ink has faded or the page discolored. Scholars have therefore spent a great deal of time and effort on manually rejoining leaves of the same original book or pamphlet, and on piecing together smaller fragments, usually as part of their research in a particular topic or literary work. Throughout the years, scholars have devoted a great deal of time to manually identify such groups of fragments, referred to as *joins*, often visiting numerous libraries for this purpose. Despite the several thousands of such joins that have already been identified by researchers, much more remains to be done (Lerner and Jerchow). Accordingly, to make the future study of the Genizah more efficient, there is an urgent need to group the fragments together and to try to reconstruct the original codices as well as possible.

Manual classification is currently the “gold standard” for finding joins. However this is not scalable and cannot be applied to the entire corpus. We suggest automatically identifying candidate joins to be verified by human experts. To this end we employ modern image-recognition tools such as local descriptors, bag-of-features representations, and discriminative metric learning techniques, as explained in Section 3 of this chapter. These techniques are modified by applying suitable preprocessing and by using task-specific key-point selection techniques. Furthermore, a bag of visual keywords approach is taken in which palaeographic samples of various script styles are

used. It can be shown that this step increases performance considerably. The results are presented in Sections 4 and 5.

In addition to the automated join-finding effort, we also study the problem of automatically deriving the script style of Genizah documents. We choose to do it in an *unsupervised* manner, in which a clustering algorithm groups the various documents, thereby separating the image sets according the script style of each image, with no a priori bias towards a particular classification scheme. Nevertheless, the resulting division is a close match to the standard taxonomy. This aspect of our work is the subject of Section 6.

Section 7 discusses related work and is followed by a brief summary of our achievements.

2. Image Processing and Physical Analysis

The images supplied by the Friedberg Genizah Project were in the format of 300–600 dpi JPEGs with arbitrarily aligned fragments placed on varying backgrounds. Although uncompressed images of higher resolution are available, we choose not to use these since the type of methods we use do not require higher resolution, and since the compression artifacts can be neglected in comparison to the deformations created to the original fragment over the centuries. An example, which is relatively clean, is shown in Figure 1(a). Many of the images, however, contain superfluous parts for our task, such as paper tags, rulers, color tables, etc. (as in Figure 5). Therefore, a necessary step in our pipeline is preprocessing of the images to separate fragments from the background and to align fragments so the rows of text are horizontal. Then the physical properties of the fragments and of the text lines are measured. Both stages are described in detail in a previous work (Wolf et al.).

2.1. Preprocessing

The goal of the preprocessing stage is to eliminate parts of the images that are irrelevant or may bias the join finding process, and to prepare the images for the representation stage.

Coarse manual alignment. In a first manual stage, the written sides of each fragment were identified. All the images were then manually rotated as necessary in multiples of 90°, resulting in alignment in the range of $[-45^\circ, 45^\circ]$ from upright. This initial rotation prevents the following auto-alignment from rotating documents upside-down. Both the identification of the written side and the coarse alignment stages are now being automated; however, the manual effort expended for the work reported here was not great.

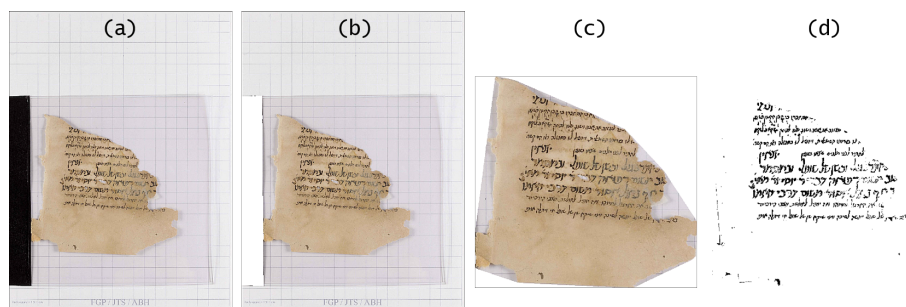


Figure 1. Example of a document from the Cairo Genizah (ENA collection). (a) The original image. (b) After removing the black folder. (c) After segmentation (using the convex hull). (d) After binarization and alignment.

Foreground segmentation. The process of separating fragments from the background in the photographs depends on the way the image was captured. At first, a machine classifier was used to identify foreground pixels based on RGB color values or HSV values. To create a region-based segmentation of the fragments, the connected components of the detected foreground pixels are marked, and the convex hull of each component is calculated. By *connected component*, we mean a contiguous region of foreground pixels; by *convex hull*, we mean the smallest possible encompassing convex (angles opening inward) polygon. Those steps retain almost all of the relevant parts of the images while excluding most of the background.

Detection and removal of non-relevant components. Labels, ruler, color swatches, and any other non-relevant components that fall in separated regions were manually removed. In some images, especially of large documents, a ruler appears adjacent to the actual fragments and is not separated by the region-segmentation process. The ruler used in the images is of a known type, so we locate it by an automated detector based on correspondence to a reference image of this ruler. The correspondence is done by employing a randomized algorithm, RANSAC (Fischler and Bolles), in combination with scale-invariant feature transform (SIFT) (Lowe) keypoint matching. The region of the detected ruler is segmented by color and removed.

Binarization. The regions detected in the foreground segmentation process are then binarized, that is, every ink pixel is assigned a value of 1 (representing black), and all other pixels are assigned a value of 0 (for white). This is done using the auto-binarization tool of the ImageXpress 9.0 package by Accusoft Pegasus. To cope with failures of the Pegasus binarization, we also binarized the images using the local threshold set at 0.9 of the local average of the 50x50 patch around each pixel. The final binarization is

the pixel-wise AND of those two binarization techniques. Pixels near the fragment boundary are set to 0. A sample result is shown in Figure 1(b). Experiments with more sophisticated binarization methods, such as Bar-Yosef et al. (2007), are ongoing.

Auto-alignment. Each region is automatically rotated so the rows (lines of text) are in the horizontal direction. This is done using a simple method, which is similar to Baird and to Srihari and Govindaraju. For each possible rotation angle we consider the ratio of black (binary value 1) to white (binary value of 0) pixels for each horizontal line. We then calculate the variance of the projection for each angle, and select the angle for which the variance is the largest.

Physical measurements. The measurements that are being used in fragment matching are characteristics of the text rows, and dimensions of the *text* bounding box (smallest rectangle containing all the text). The number of text rows, height of the rows and the spaces between the rows are calculated automatically using the projection profile of the fragment (the proportion of black in each row of pixels). The text rows themselves are localized at the maxima points of these projections. In addition, the minimal-area bounding box of each fragment is computed. Note that this bounding box need not be axis-aligned.

3. Image Handwriting Representation

We decided to employ a general framework for image representation that has been shown to excel in domains far removed from document processing, namely, a method based on a bag of visual keywords (Dance et al.; Lazebnik, Schmid, and Ponce). The “signature” of a leaf is based on descriptors collected from local patches in its fragments, centered around key visual locations, called “keypoints”. Such methods follow this pipeline: first, keypoints in the image are localized by examining the image locations that contain most visual information. In our case, the pixels of the letters themselves are good candidates for keypoints, while the background pixels are less informative. Next, the local appearance at each such location is encoded as a vector. The entire image is represented by the obtained set of vectors, which in turn is represented as a single vector. This last encoding is based on obtaining a “dictionary” containing representative prototypes of visual keywords and counting, for each image, the frequency of visual keywords that resemble each prototype appearing in the dictionary.

3.1. Keypoint Detection

We detect the image keypoints using the fact that, in Hebrew writing, letters are usually separated. We start by calculating the connected components (CCs) of the binarized images. To filter broken letter parts and dark patches arising from stains and border artifacts, we compare the size of the CC to the height of the lines, which is estimated

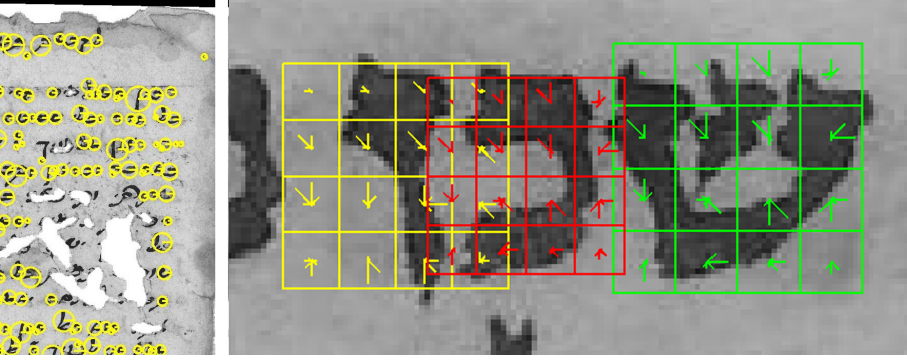


Figure 2. (a) Keypoint detection methods using the proposed CC method. (b) SIFT descriptors of three neighboring detected keypoints.

in a similar fashion to the alignment stage of the previous section. For encoding, each keypoint requires a scale, which is taken as the maximum dimension of the associated CC. Figure 2(a) shows the keypoints found using the SIFT and CC detectors.

3.2. Local Descriptors

Each keypoint is described by a descriptor vector. After experimenting with a large variety of such descriptors, the most popular descriptor, the scale-invariant feature transform (SIFT), was chosen for its accuracy. SIFT (Lowe) encodes histograms of gradients in the image. Figure 2(b) illustrates the application of SIFT to one fragment.

3.3. Dictionary Creation and Vectorization

Bag-of-visual-keyword techniques (Dance et al.) rely on a dictionary that contains a representative selection of descriptors obtained on various interest points. To this end, we first set aside a small dataset of 500 documents. We detect keypoints in those documents and subsample a large collection of 100,000 descriptors. These are then clustered by the k -means algorithm to obtain a dictionary of varying sizes.² The result is a set of prominent prototypes or “visual keywords”; see Figure 3.

² Clustering algorithms (in machine learning parlance) assign input samples to homogenous groups that are distinctive from each other. The k -means algorithm is one of the simplest such algorithms. After an initialization stage, it repeats two steps multiple times: first, each sample is assigned to a cluster based on its distance to all cluster centers, and second, each cluster center is updated to be the mean vector value of all points that were assigned to this cluster.

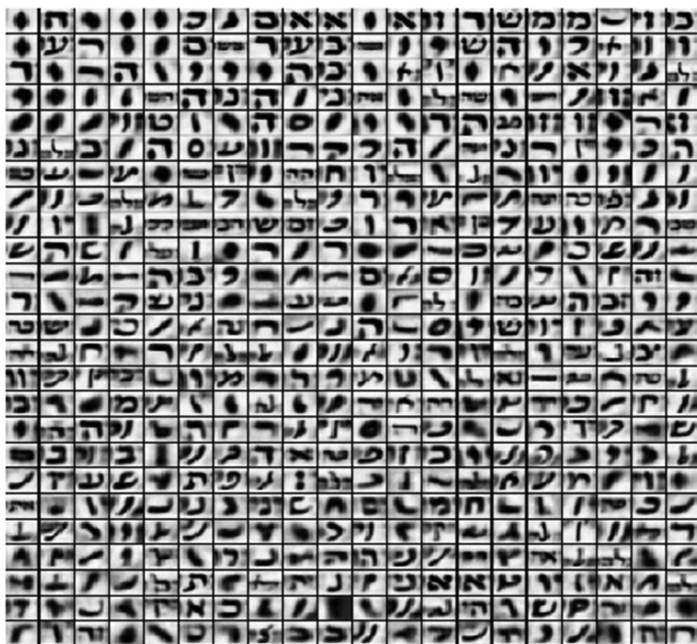


Figure 3. Cluster “centers” obtained by taking the average image of all images in each cluster. By clustering (grouping into homogenous groups) a large collection of visual descriptors obtained from random images, a set of distinctive visual keywords or prototypes, referred to as a “dictionary”, is formed. Note that the cluster centers look blurry, since they stem from averaging multiple images.

Given a dictionary, we employ either histogram-based methods or distance-based methods to encode each manuscript leaf as a vector. In histogram-type vectorization methods (Dance et al.), for each cluster-center in the dictionary, one counts the number of leaf descriptors (in the encoded image) closest to it. The result is a histogram of the descriptors in the encoded leaf with as many bins as the size of the dictionary. To account for the variability in fragment sizes we normalize the histogram vector to sum to 1, that is, we divide the histogram vector by its sum. Another alternative is to normalize each vector so that its Euclidean norm is 1.³

Distance-based representation techniques (Serre, Wolf, and Poggio) are based on computing the minimum distance to all descriptors of the given leaf for each prototype in the dictionary.

³ The Euclidean, or L2, norm is the square-root of the sum of the squares of each dimension. An L1 norm is just the sum of absolute values.

Script style	Number of samples	Page list
Square and Semi-Square Oriental	51 images	I-1 – I-51
Semi-cursive Oriental	61 images	I-52 – I-112
Yemenite	42 images	I-113 – I-154
Square Spanish	62 images	II-1 – II-62
Semi-cursive Spanish	99 images	II-63 – II-161
Cursive Spanish	48 images	II-162 – II-209

Table 1. The images of document samples used to obtain dictionaries for each script style were taken from the indicated pages of the palaeography volumes I (Beit-Arie, Engel, and Yardeni 1987) and II (Beit-Arie, Engel, and Yardeni 2002).

3.4. Employing Palaeographic Data in the Representation

The resulting representation transforms each image into a vector based on a global dictionary, in which prototypes from various script types are mixed together. Since most Genizah joins contain just one script style, it might be helpful to have multiple dictionaries, each corresponding to one script style. To obtain such dictionaries we apply the procedure described in Section 3.3 repeatedly, each time on documents of the same script.

Sample documents for each style were extracted from the pages of the medieval Hebrew script specimen volumes by Beit-Arie, Engel, and Yardeni (1987, 2002). These volumes contain many examples of medieval Hebrew manuscripts whose provenances are known, and serve as an important tool in Hebrew palaeography. High-quality sample pages of manuscripts are printed side-by-side with professionally-drawn sample letters of the alphabet, based on exemplars from the manuscript. Note that we use the images of the original documents and not the individually hand-drawn letter samples.

The groups of script styles obtained from this collection, and the corresponding page numbers of the collection are listed in Table 1. As can be seen, the major script styles are square script, semi-cursive script, and cursive script. The geographic location influences script style, so we extracted Oriental, Yemenite, and Spanish script groups from the same source.

4. Finding Joins

To determine whether two fragments originate from the same manuscript, we compare their vector representations. The comparison can be performed in several ways and it is often beneficial to combine multiple methods.

4.1. Similarity Inference

Focusing on just one representation, each leaf is represented by one vector, for example, by the L2-normalized histogram of keypoint types. For every pair of leaves, we need to determine whether they are from the same join or not. Ideally, we would have a similarity function that would return a high value when two leaves are from the same join, and a low value otherwise. In this ideal case, a threshold value of the similarity function provides a decision cutoff value.

The basic similarity score is obtained by considering, for every two vectors p and q , the similarity derived from their Euclidean distance $e^{-\|p-q\|}$.

In our work we also employ learned similarities. Tailoring similarity measures to available training data by applying learning techniques is gaining popularity. Here, the similarity is to be learned from pairs of samples that are known to belong to the same join or not, and we choose to use a similarity that has been shown to be extremely successful in face-recognition work.

The One Shot Similarity (OSS) (Wolf, Hassner, and Taigman 2008, 2009) is a similarity learning technique designed for the same/not-same problem. Given two vectors p and q , their OSS score is computed by considering a training set of background sample vectors A . This set of vectors contains examples of items different from either p and q (that is, they do not belong in the same class as neither p or q). Note, however, that these training samples are otherwise unlabeled. In our experiments we take the set A to be one split out of the nine splits used for training at each iteration (see Section 4.3).

A measure of the similarity of p and q is then obtained as follows. First, a discriminative model is learned⁴ with p as a single positive example and A as a set of negative examples. This model is then used to classify the second vector, q , and obtain a classification score. The nature of this score depends on the particular classifier used. We employ a Linear Discriminant Analysis (LDA) classifier, and the score is the signed distance of q from the decision boundary learned using p (positive example) and A (negative examples). A second such score is then obtained by repeating the same process with the roles of p and q switched: this time, a model learned with q as the positive example is used to classify p , thus obtaining a second classification score. The final OSS is the sum of these two scores.

⁴ Classifiers or learned discriminative models (in machine learning terminology) are functions whose parameters are fit in a way that they predict the class of a given input. Typically, training samples are given that are divided into two sets—a positive set and a negative set. Learning then takes place by computing the function parameters that would assign a positive or a negative label to every training sample similarly to the given labels. In this paper, we use two classification algorithms: Linear Discriminant Analysis (LDA), which is a learning method that assumes Gaussian conditional density models, and linear Support Vector Machine (SVM), a classifier that strives to separate the positive samples from the negative ones as much as possible.

4.2. Classification and Combinations of Features

For the recognition of joins we need to convert the similarity values of Section 4.1 to a decision value. Moreover, it is beneficial to combine several similarities. For both these tasks we employ linear support vector machines (SVM), with fixed parameter value $C = 1$, as was done in Wolf, Hassner, and Taigman (2008) and Wolf, Bileschi, and Meyers (2006).

In the case of one-similarity, the similarity is fed to SVM as a one-dimensional vector and training is performed on all training examples. In this case, SVM just scales the similarities and determines a threshold for classification.

To combine several similarities together we use the SVM output (signed distance from dividing hyperplane) obtained separately from each similarity and construct a vector. This vector is then fed to another SVM. The value output by the last classifier is our final classification score. This method of combining classifier output is called “stacking” (Wolpert).

4.3. The Genizah Benchmark

To evaluate the quality of our join-finding efforts, we constructed a comprehensive benchmark. Our benchmark, modeled after the LFW face recognition benchmark (Huang et al.), consists of 31,315 leaves, all from the New York (ENA), Paris (AIU), and Jerusalem (JNUL) collections.

The benchmark consists of ten equally sized sets. Each contains 1000 positive pairs of images taken from the same joins, and 2000 negative (non-join) pairs. Care is taken to ensure that no known join appears in more than one set, and that the number of positive pairs taken from one join does not exceed 20.

The ROC (receiver operating characteristic) curve is an accepted form of measuring classification success. It is a graph (see Figure 4) in which the trade-off between false positive (type I error) results and the recall (true positive) rate is displayed. One would like to obtain perfect recall (identifying all joins) making no false-positive errors, that is, without identifying non-joins as joins. However, in reality the task is challenging and therefore a certain number of false detections is expected for reaching high levels of recall.

To report results, the classification process is repeated 10 times. In each iteration, nine sets are taken as training, and the results are evaluated on the tenth set. Results are reported by constructing an ROC curve for all splits together (the outcome value for each pair is computed when this pair is a testing pair), by computing statistics of the ROC curve (area under curve, equal error rate, and true positive rate at a certain low false positive rate) and by recording average recognition rates for the 10 splits.

The most interesting statistic from the practical point of view is the recall at a low-false positive rate. Since there are many theoretical join candidates in the Genizah and since human verification effort is limited, any practical join-finding system should mark non-joins as joins only for a small percentage of these candidates.

4.4. Benchmark Results

We compare the performance of several methods, each based on a separate source of information. Not surprisingly, combining these methods yields the best results.

Subject classification. Over 95% of the digitized Genizah documents have already been manually classified by subject matter. The classification contains categories such as “Biblical”, “Correspondence”, “Liturgy”, “Arabic tafsir”, “Aramaic translation”, and more. A similarity of -1 is assigned to pairs of documents with incompatible classifications. A score of +1 is given if the classifications are compatible, and a score of 0 when compatibility cannot be determined.

Physical measurements. Eight measurements are considered: number of lines, average line height, standard deviation of line height, average space between lines, standard deviation of interline space, minimal bounding box width, minimal bounding box height, and area of the minimal bounding box. Each one of these measurements is hardly discriminative; however, combined together, they are able to discriminate pretty reliably between joins and random pairs, although not as well as the handwriting approach below.

Handwriting. The handwriting is represented using the bag of visual keywords approach described above. With a global dictionary, the best performing method uses the One-Shot-Similarity (OSS) of Section 4.1.

Multiple script-style dictionaries. The OSS scores obtained from the various dictionaries described in Section 3.4 are combined using the stacking technique of Section 4.2. This method provides a noticeable improvement over the single-dictionary method.

Combined methods. In addition, we combine the handwriting-based scores (single or multiple dictionaries) with the physical score and with the subject-classification score.

The results are summarized in Table 2. It can be seen that the best method, the one that combines the multiple script-style dictionaries with the physical measurements and the subject classification, obtains a recall rate of up to 84.5% at a false-positive rate of 0.1%. The obtained ROC curves are depicted in Figure 4(a). While some of the improvements seem incremental, they actually make a significant difference in the low false-positive region (Figure 4(b)).

Method	Area under ROC	Equal error rate	Mean success \pm standard error	TP rate at FP rate of 0.001
Subject classification	0.7932	0.3081	0.4935 ± 0.0042	0
Physical measurements	0.9033	0.1843	0.8483 ± 0.0034	0.3596
Single dictionary	0.9557	0.0918	0.9374 ± 0.0048	0.7600
Single dictionary + physical	0.9785	0.0627	0.9566 ± 0.0028	0.8116
Multiple script-style dictionaries	0.9805	0.0564	0.9596 ± 0.0029	0.8053
Multiple dictionaries + physical	0.9830	0.0524	0.9625 ± 0.0028	0.8229
Multiple + physical + subject	0.9888	0.0430	0.9680 ± 0.0024	0.8451

Table 2. Results obtained for various similarity measures and combinations thereof. See text for the description of each method.

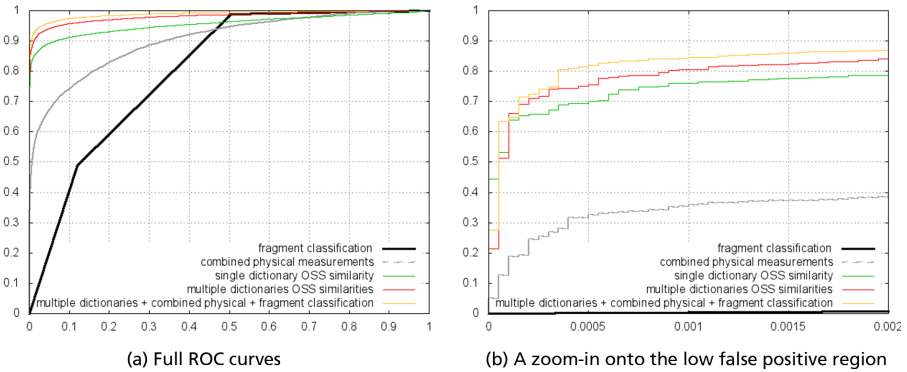


Figure 4. ROC curves (true positive rate vs. false positive rate) averaged over 10 folds. The plots compare the results obtained for the subject classification criterion, physical measurements, a single global dictionary, multiple script-type dictionaries, and the multiple dictionary approach combined with physical measurements and subject classification.

5. Newly-Found Joins

We have conducted three sets of examinations to evaluate the value of our system in finding joins beyond the settings of the benchmark.

5.1. A Small Benchmark

A set of experiments was performed on an initial benchmark we created for a preliminary work (Wolf et al.). This benchmark was much smaller and contained ten splits each containing 196 positive pairs and 784 negative ones. All images were taken from the ENA and AIU collections. As mentioned above, the negative pairs we work with are not necessarily negative. This does not affect the numerical results much, since the fraction of joins is overall low; however, it implies that there may exist unknown joins in the set of leaves that are currently available to us.

We applied our classification technique to all possible pairs of leaves and then looked at the 30 leaf pairs that were not known to be joins, but which received the highest matching scores. The resulting pairs were submitted to a human expert for validation. The manual labor involved was about 2.5 hours. Eighty percent of the newly detected join candidates were actual joins. Seventeen percent are not joins and one pair could not readily be determined.

5.2. The Geneva Collection

We applied our system to the task of locating joins with the recently recovered Geneva collection. The search for joins using our tools was pretty efficient, with about 30% of the top 100 matches returned turning out to be actual joins. Figure 5 shows a variety of previously-unknown joins proposed by our method. Example (a) consists of two leaves from the same copy of the Mishnah, written on vellum in Hebrew in a square script. The texts are from different tractates of *Order Zeraim*. The left page is from the Geneva collection and the right one from the small collection of the Jewish National and University Library (JNUL). Other leaves from the same manuscript are in Oxford and Cambridge.⁵ Example (b) shows fragments from a codex of the Bible, both from the book of Exodus (Hebrew, square script, on vellum), one from Geneva and the other from the Jewish Theological Seminary (JTS) in New York, part of a batch of 69 fragments from various biblical manuscripts (partially vocalized and with cantillation signs). Such codices are written using a very rigid set of typographic rules, and the identification of such joins based on handwriting is considered extremely challenging. Example (c) is in alternating Hebrew and Aramaic (*Targum*, square script), one page from Geneva and

⁵ It turns out that this specific automatically-proposed join has already been discovered and is documented in the very recent Geneva catalogue (Rosenthal), and in the forthcoming Sussmann Catalog.

Range	Strong join	Weak join	Total join	Excluding empty
1–2000	17.05%	6.95%	24.00%	44.8%
5791–8790	7.16%	6.20%	13.37%	18.0%

Table 3. The percentile of verified new joins out of the candidate joins suggested by our system.

the other from the New York JTS collection. Example (d) shows a join of two leaves of Hebrew liturgical supplications from Geneva and from Pennsylvania, in rabbinic script. Example (e) is from a book of precepts by Saadiah ben Joseph al-Fayyumi, a lost halakhic work by the 10th century gaon. The left page is from the Geneva collection and the right one from JTS. The language is Judeo-Arabic, and the text is written in a square oriental script on vellum. This is a good example of how joins can help identify new fragments from lost works. Once one member of a pair is identified correctly, the identification of the second one is self-determined. Example (f) is from a responsum in Hebrew (rabbinic script). Both leaves are from the Alliance Israelite Universelle Library in Paris, but they are catalogued under different shelfmarks.

5.3. Between Collections

A third set of join-seeking efforts was conducted on all between-collection pairs of fragments unknown to be joins in the ENA, AIU, and JNUL collections, as well as in smaller European collections of mixed quality.

Note that inter-collection joins are harder for humans to find, and are more challenging and rare. The top scoring 9,000 pairs were extracted. After further analysis of catalogue information some additional known pairs were removed resulting in 8,790 pairs. The first 2,000 pairs and the last 3,000 fragments of this list were studied. The results are given in Table 3. It separates between “strong” joins, meaning same scribe and same manuscript, and “weak” joins—a join between different manuscripts that seem to be written by the same scribe. In contrast to strong joins, the certainty of a weak join coming from the same document is doubtful, and in many cases should be examined carefully again by an expert. In any event, a weak join represents a good candidate for fragments written by the same scribe, and as such it is considered a success.

As can be seen, 24% of the top discoveries are true joins, mostly strong. More than 13% of the 6th, 7th, and 8th thousands of matches are validated joins. At least half of those are strong joins. Going over the examples it became apparent that many of the proposed joins were artifacts caused by normalized vectors arising from empty documents. This was to be expected, since the benchmark that was used to develop the join-discovery

tool was not designed to handle blank documents. After the removal of 49 empty fragments and all their discovered joins, the recognition rates grew considerably.

6. Unsupervised Grouping by Script Style

As we have found, the most distinguishing visual information between the fragments arises from the handwriting. The search for joins focuses on minute differences that exist between various scribes. We now turn our attention into grouping the documents by a much coarser distinction: the one between script styles.

We sample 300 leaves from the Genizah collection that have been classified into one of 12 script styles: “Square Ashkenazi”, “Square Italian”, “Semi-cursive Oriental”, “Square Oriental”, “Cursive Oriental”, “Semi-cursive Spanish”, “Square Spanish”, “Cursive Spanish”, “Semi-cursive Yemenite”, “Square Yemenite”, “Square North-African”, “Cursive North-African”. We then attempt to group the leaves automatically, a process called “clustering”.

We found that conventional clustering algorithms such as k -means work poorly for separating the documents into script-styles. Indeed, k -means focuses on clusters of similar sizes, and might produce unintuitive results for data that is not distributed homogeneously in the parameter space.

We therefore employed the following method that was developed in order to deal with an unknown number of clusters, variability in cluster size, and inhomogeneous data.

First, each leaf is represented as a vector using the bag of visual keyword approach and a single global dictionary. Multiple dictionaries would not be appropriate here, since we would like to obtain the script styles from the data, and not impose it on the representation.

Recall that the vector representing each leaf contains visual “keyword” frequencies. To eliminate noise and remove spurious correlations between documents, we focus on the most prominent keywords for each document. This is done by replacing each keyword frequency that is less than half of the maximal frequency by 0.

In the next step, we build a graph in which every leaf is a node, and an edge exists between two nodes if the correlation between their modified vectors is above 0.5. The connected components of this graph are taken as the initial clusters. Connected components that contain single points are referred to below as “singletons” and are considered unclustered.

We then refine these clusters by iterating, until convergence, two alternating steps. In the first step, pairs of clusters for which the distances between each cluster’s points and the cluster mean point are similar to the distances between the two clusters are merged. In the second step, singletons are assigned to clusters if their distance to the

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 8	cluster 9	cluster 10	cluster 11	cluster 12	cluster 13	cluster 14	cluster 15	cluster 16	cluster 17	cluster 18	unclustered
Square Ashkenazi	0.00	0.00	0.00	0.33	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
Square Italian	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Semi-cursive Oriental	0.00	1.00	1.00	0.67	0.00	0.00	0.20	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15
Square Oriental	0.00	0.00	0.00	0.00	0.64	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18
Cursive Oriental	0.00	0.00	0.00	0.00	0.04	0.00	0.80	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.03
Semi-cursive Spanish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
Square Spanish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15
Cursive Spanish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.15
Semi-cursive Yemenite	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Square Yemenite	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.06
Square North-African	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.09
Cursive North-African	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.71	1.00	0.00

Table 4. A confusion matrix that shows that frequency within each obtained cluster of each script style. For example, the first cluster is composed entirely out of leaves of square Italian script style, while the forth cluster is one-third square Ashkenazi and two-thirds Semi-cursive Oriental.

closest cluster is not larger than three times the standard deviation of distances within that cluster.

After convergence, this procedure yields 18 clusters and 34 singletons. The clusters are pretty homogenous with regard to script style: 93% of the documents are clustered within clusters in which their script-style is the most frequent script-style; 7% are clustered in clusters in which they belong to the minority.

The distribution of documents of various script styles among the 18 clusters is shown in the confusion matrix presented in Table 4. Each row of this matrix corresponds to one script style, and each column to one cluster.

Figure 6 shows samples from representative clusters. As can be seen, confusion is often a result of script styles that are superficially similar. Naturally a more detailed analysis of individual letters would lead to more accurate results; however, this requires accurate optical character recognition, which is beyond the current state of the art for the vast majority of Genizah images.

7. Related Work

7.1. Writer Identification

A related task to that of join finding is the task of scribe identification, in which the goal is to identify the writer by morphological characteristics of a writer’s handwriting. Since historical documents are often incomplete and noisy, preprocessing is often applied to separate the background and to remove noise (Bres, Eglin, and Volpillac-Augur; Leedham et al.). Latin letters are typically connected, unlike Hebrew ones which are usually only sporadically connected. Efforts were thus expended on designing segmentation algorithms to disconnect letters and facilitate identification (Casey and

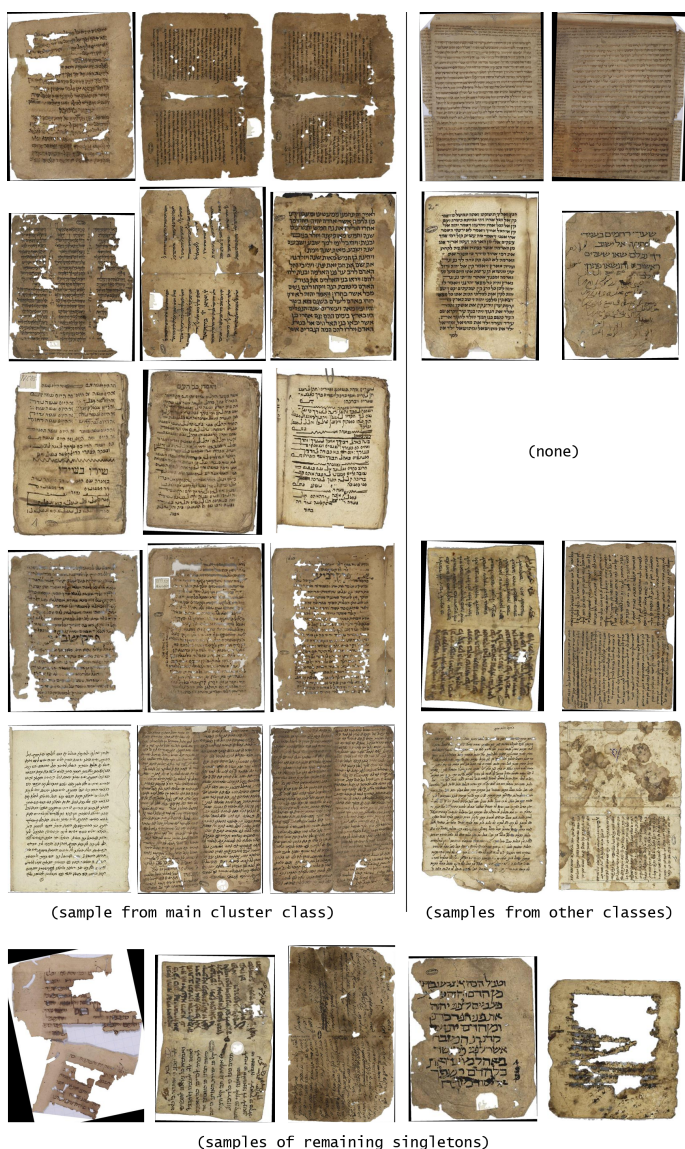


Figure 6. Each row shows samples from a single cluster. The left three samples are from the main script style of this cluster, while the two samples to the right are samples of the same cluster that belong to other script styles. Shown, from top to bottom are clusters 4, 5, 6, 8, 17. Also shown (last row) are samples of singletons, i.e., documents that were not assigned to any of the clusters.

Lecolinet). The identification itself is done either by means of local features or by global statistics. Most recent approaches are of the first type and identify the writer using letter- or grapheme-based methods, which use textual feature matching (Panagopoulos et al.; Bensefia, Paquet, and Heutte). The work of Bres, Eglin, and Volpilhac-Auger (2006) uses text-independent statistical features, while other efforts combine both local and global statistics (Bulacu and Schomaker 2007a; Dinstein and Shapira).

Interestingly, there is a specialization to individual languages, employing language-specific letter structure and morphological characteristics (Bulacu and Schomaker 2007a; Panagopoulos et al.; Dinstein and Shapira). In our work, we rely on the separation of Hebrew characters by employing a keypoint detection method that relies on connected components in the thresholded images.

Most of the abovementioned works identify the writer of the document from a list of known authors. Here, we focus on finding join candidates, and do not assume a labeled training set for each join. Still, since writers are usually unknown (in the absence of a colophon or signature), and since joins are the common way to catalog Genizah documents, we focused on this task. Note that the handwriting techniques we use are not entirely suitable for distinguishing between different works of the same writer. However, additional data, such as text or topic identification, page size and number of lines, as used in Section 4, can help distinguish different works by the same writer.

7.2. Digital Palaeography

Palaeographers traditionally use a mix of qualitative and quantitative features to distinguish hands (Mallon). Early uses of image analysis and processing for palaeographic research include the work of Founder and Vienot, Sirat, and Dinstein and Shapira; Plamondon and Lorette survey other early work. Quantitative aspects can be measured by automated means and the results can be subjected to computer analysis and to automated clustering techniques (Ciula; Aussems; Aioli and Ciula). Features amenable to automatization, including texture (Said, Tan and Baker; Bulacu and Schomaker 2007b), angularities (Bulacu, Schomaker, and Vuurpijl), and others (Aussems and Brink) have been suggested. Concavity, moments, and other features have been used to correctly classify selected Hebrew letters by writer (Bar-Yosef et al. 2004, 2007). What distinguishes our work is that we are using generic image features for this purpose.

8. Conclusion

We have presented a framework for identifying joins in Genizah fragments, which has already provided results of value to Genizah researchers by identifying heretofore unknown joins. We have shown how handwriting data, especially when combined with

prior knowledge of script styles, physical measurements, and subject classification, can produce a reliable system.

Through our semi-automated efforts approximately 1000 new joins have been identified. Given that the overall number of joins found in over a century of Genizah research and by hundreds of researchers is only a few thousand, our system has proved its scalability and value. The main limiting factor in finding more joins is the short supply of human experts. We hope to alleviate this constraint by making our join candidates available over the internet to the Genizah research community.

We also explored the grouping of Genizah documents in a top-down manner, and have shown that, when the heterogeneous nature of the data set is accounted for, the palaeographic information emerges as the most visually prominent characteristic.

The methods presented here are applicable to other corpora as well. Many archives hold large unstructured sets of handwritten forms, letters, or other documents. The same technology could provide meta-data and enable queries based on similarity, and automatic grouping of the documents. The information employed is complementary to that obtained by Optical Character Recognition (OCR) systems, and would remain so even were the accuracy of the OCR systems to increase substantially. Note that although we did not focus on Latin scripts, the method is suitable to such scripts as well, with relatively straightforward adaptations to the keypoint mechanisms.

Bibliography

- Aiolfi, Fabio and Arianna Ciula. "A Case Study on the System for Paleographic Inspections (SPI): Challenges and New Developments." *Proceeding of the 2009 Conference on Computational Intelligence and Bioengineering*. Amsterdam, IOS Press, 2009. 53–66.
- Aussems, Mark. *Christine de Pizan and the Scribal Fingerprint – A Quantitative Approach to Manuscript Studies*. Master's thesis. Utrecht, 2006.
- Aussems, Mark and Axel Brink. "Digital Palaeography." *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Eds. Malte Rehbein, Patrick Sahle, and Torsten Schaßan. Schriftenreihe des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 293–308.
- Baird, K.S. "Anatomy of a Versatile Page Reader." AT&T Bell Lab., Murray Hill, NJ. *Proceedings of the IEEE* 80.7 (1992): 1059–1065.
- Bar-Yosef et al. "Classification of Hebrew Calligraphic Handwriting Styles: Preliminary Results." *DIAL '04: Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL '04)*. Washington, DC, USA: IEEE Computer Society, 2004. 299–305.
- Bar-Yosef et al. "Binarization, Character Extraction, and Writer Identification of Historical Hebrew Calligraphy Documents." *International Journal on Document Analysis and Recognition* 9 (2007): 89–99.

- Beit-Arie, Malachi, Edna Engel, and Ada Yardeni. *Specimens of Mediaeval Hebrew Scripts, Volume 1: Oriental and Yemenite Scripts (in Hebrew)*. Jerusalem: The Israel Academy of Sciences and Humanities, 1987.
- Beit-Arie, Malachi, Edna Engel, and Ada Yardeni. *Specimens of Mediaeval Hebrew Scripts, Volume 2: Sefardic Script (in Hebrew)*. Jerusalem: The Israel Academy of Sciences and Humanities, 2002.
- Bensefia, Ameer, Thierry Paquet, and Laurent Heutte. "Information Retrieval Based Writer Identification." *Seventh International Conference on Document Analysis and Recognition, Volume 2*. Mont-Saint-Aignan: Laboratoire Perception Systèmes Information, UFR des Sciences, Université de Rouen, 2003. 946–950.
- Bres, Stephane, Veronique Eglin, and Catherine Volpillac-Auger. "Evaluation of Handwriting Similarities Using Hermite Transform." *Tenth International Workshop on Frontiers in Handwriting Recognition*. Ed. Guy Lorette La Baule (France): Suvisoft, 2006.
- Bulacu, Marius L. and Lambert R.B. Schomaker. "Automatic Handwriting Identification on Medieval Documents." *14th International Conference on Image Analysis and Processing, Groningen*: Univ. of Groningen, 2007. 279–284.
- Bulacu, Marius L. and Lambert R.B. Schomaker. "Text-Independent Writer Identification and Verification Using Textural and Allographic Features." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (2007): 701–717.
- Bulacu, Marius L., Lambert R.B. Schomaker, and Louis Vuurpijl. "Writer Identification Using Edge-Based Directional Features." *Proceedings of the Seventh International Conference on Document Analysis and Recognition*. Washington, DC: IEEE Computer Society, 2003. 937–941.
- Casey, Richard G. and Eric Lecolinet. "A Survey of Methods and Strategies in Character Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18 (1996): 690–706.
- Ciula, Arianna. "Digital Palaeography: Using the Digital Representation of Medieval Script to Support Palaeographic Analysis." *Digital Medievalist* 1.1 (2005).
<<http://www.digitalmedievalist.org/journal/1.1/ciula/>>.
- Dance, Chris, Jutta Willamowski, Lixin Fan, Cedric Bray and Gabriela Csurka. "Visual Categorization with Bags of Keypoints." *ECCV Workshop on Statistical Learning in Computer Vision*. 2004. 1–22.
- Dinstein, Its'hak and Yaacov Shapira. "Ancient Hebraic Handwriting Identification with Run-length Histograms." *IEEE Transactions on Systems, Man and Cybernetics* 12 (1982): 405–409.
- Fischler, Martin A. and Robert C. Bolles. "Random Sample Consensus: a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography." *Communication of the ACM* 24 (1981): 381–395.
- Fournier, Jean-Marc and Jean-Charles Vienot. "Fourier Transform Holograms used as Matched Filters in Hebraic Paleography." *Israel Journal of Technology* (1971): 281–287.
- Huang, Gary B., Manu Ramesh, Tamara Berg, and Erik Learned-Miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." University of Massachusetts, Technical Report 07-49, 2007.

- Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006. 2169–2178.
- Leedham, Graham, Saket Varma, Anish Patankar, and Venu Govindarayu. "Separating Text and Background in Degraded Document Images; A Comparison of Global Thresholding Techniques for Multi-Stage Thresholding." *Eighth International Workshop on Frontiers in Handwriting Recognition*, 2002. 244.
- Lerner, Heidi G. and Seth Jerchow. "The Penn/Cambridge Genizah Fragment Project: Issues in Description, Access, and Reunification." *Cataloging & Classification Quarterly* 42 (2006): 21–39.
- Lowe, David G. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision* 60 (2004): 91–110.
- Mallon, Jean. *Paleographie Romaine*. Madrid: Consejo Superior de Investigaciones Científicas, Instituto Antonio de Nebrija, de Filología, 1952.
- Panagopoulos, Michail et al. "Automatic Writer Identification of Ancient Greek Inscriptions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (2009): 1404–1414.
- Plamondon, Rjean and Guy Lorette. "Automatic Signature Verification and Writer Identification – the State of the Art." *Pattern Recognition* 22 (1989): 107–131.
- Reif, Stefan C. *A Jewish Archive from Old Cairo: The History of Cambridge University's Genizah Collection*. Richmond (England): Curzon Press, 2000.
- Rosenthal, David. *The Cairo Genizah Collection in Geneva: Catalogue and Studies*. Jerusalem: Magnes Press, 2010.
- Said, Huwida E. S., Tienniu N. Tan, and Keith D. Baker. "Personal Identification based on Handwriting." *Pattern Recognition* 33 (2000): 149–160.
- Serre, Thomas, Lior Wolf, and Tomaso Poggio. "Object Recognition with Features Inspired by Visual Cortex." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2 (2005): 994–1000.
- Sirat, Colette. *L'examen des écritures: l'oeil et la machine. Essai de méthodologie*. PhD thesis, Paris: Editions du Centre National de la Recherche Scientifique, 1981.
- Srihari, Sargur N. and Venugopal Govindaraju. "Analysis of Textual Images Using the Hough Transform." *Machine Vision and Applications* 2 (1989): 141–153.
- Wolf, Lior, Tal Hassner, and Yaniv Taigman. "The One-Shot Similarity Kernel." *IEEE International Conference on Computer Vision (ICCV)*. 2009. 897–902.
- Wolf, Lior et al. "Automatically Identifying Join Candidates in the Cairo Genizah." *Post ICCV workshop on eHeritage and Digital Art Preservation*. 2009.
- Wolf, Lior, Tal Hassner, and Yaniv Taigman. "Descriptor Based Methods in the Wild." *Faces in Real-Life Images Workshop in ECCV*. 2008. <<http://hal.inria.fr/REALFACES2008/en>>.
- Wolf, Lior, Stan Bileschi, and Ethan Meyers. "Perception Strategies in Hierarchical Vision Systems." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006. 2153–2160.
- Wolpert, David H. "Stacked Generalization." *Neural Networks* 5 (1992): 241–259.

Zum Einsatz von Synchrotronstrahlung bei der Wiedergewinnung gelöschter Texte in Palimpsesten mittels Röntgenfluoreszenz

Daniel Deckers, Leif Glaser

Zusammenfassung

Gelöschte Schrift in mittelalterlichen Palimpsesthandschriften kann durch zweidimensionale Abbildung geringer Spuren der in den Pergamentblättern enthaltenen Metalle lesbar gemacht werden. Die bisherigen Versuche greifen auf Röntgenstrahlung aus einem Elektronen-Speicherring zurück, sind aber so zeitintensiv, dass nur wenige einzelne Blätter probeweise untersucht werden konnten. Anhand des Versuchsaufbaus und der bisherigen Ergebnisse werden die Perspektiven und künftigen Entwicklungsmöglichkeiten auf dem Weg zu einem regelmäßig und für ganze Handschriften einsetzbaren Verfahren aufgezeigt.

Abstract

Two-dimensional mapping of trace metals from parchment palimpsest manuscripts facilitates the reading of the erased original writing. Experiments so far rely on high-flux storage ring X-ray radiation but may take up to several days per leaf, which is practically limiting the process to sample examination of single leaves only. This chapter describes the experimental setup and results obtained so far and outlines how future development might lead to a productive workflow that would eventually allow the application of this technology to entire manuscripts.

1. Einleitung

Dieser Arbeitsbericht beschreibt Aspekte der Untersuchung von Palimpsesten in einer fortdauernden Kooperation zwischen Paläographen, Physikern und Handschriftenbibliothekaren.¹ Bei bisher drei Messzeiten (jeweils knapp einwöchigen Perioden, während

¹ Partnerinstitutionen sind das Hamburger Synchrotronstrahlungslabor (HASYLAB) am Deutschen Elektronensynchrotron (DESY), das Zentrum für Handschriften- und Textforschung »Teuchos« am Institut für Griechische und Lateinische Philologie der Universität Hamburg sowie die Universitätsbibliothek Leipzig; beteiligt sind K. Appel, G. Falkenberg, L. Glaser (HASYLAB), Chr. Mackert (UB Leipzig), Chr. Brockmann, D. Harlfinger und D. Deckers (Teuchos).

deren die Strahlenquelle genutzt werden konnte) wurden griechische Palimpseste untersucht, d.h. hier mittelalterliche Pergamenthandschriften, die wiederverwendete Blätter mit einer gelöschten griechischsprachigen Erstbeschriftung enthalten. Nicht im Detail eingegangen wird auf die im 19. Jh. verbreiteten Methoden zur chemischen Wiedergewinnung der Texte (meist alles andere als schädigungsfrei) oder auf die seit Anfang des 20. Jh. eingesetzten fotografischen Verfahren einschließlich ihrer in den letzten Jahren Verbreitung findenden digitalen Varianten.²

Bei den bisherigen Untersuchungen standen grundsätzliche Erkenntnisse zu Anwendungs- und Optimierungsmöglichkeiten des Verfahrens im Vordergrund; entsprechend wird hier vor allem auf die Grundlagen und auf praktische Aspekte eingegangen, wobei weder technische Details noch weitergehende inhaltliche Fragen zu den Einzelfällen Berücksichtigung finden. Beschrieben werden die technischen Grundlagen, der Versuchsaufbau, die Ergebnisse (auch im Vergleich zu optischen Untersuchungsmethoden) und die Perspektiven für den künftigen Einsatz des Verfahrens.

Die untersuchten Handschriften wurden von der Universitätsbibliothek Leipzig zur Verfügung gestellt.

2. Technischer Hintergrund

Um Beschreibstoff weiterzunutzen, wurden im Mittelalter bereits beschriebene Blätter aus Handschriften, die beschädigt waren oder nicht mehr benötigt wurden, chemisch oder mechanisch gelöscht und einer Wiederverwendung zugeführt. Der Wert der enthaltenen unteren Texte für die Handschriftenforschung ergibt sich schon daraus, dass wir es dabei gerade mit Material zu tun haben, das nicht als erhaltenswert angesehen wurde, und sei es nur aus materiellen Gründen wie z.B. durch Beschädigung. Im Einzelfall ist der Wert für die Editorik sehr hoch, wenn es sich um sonst nicht oder nur in anderen (insbesondere späteren) Varianten überlieferte Texte handelt, wie z.B. im Falle des berühmten lateinischen *De republica*-Palimpsests.

Spuren der Erstbeschriftung in Palimpsesten können einerseits noch vorhandene Bestandteile der Tinte sein, andererseits bei der Erstbeschriftung verursachte Veränderungen im Beschreibstoff. Eine Untersuchung mit (heute üblicherweise digitalen) fotografischen Verfahren bietet sich an, wenn Spuren der ersten Tinte noch sichtbar sind (digitale Farbaufnahmen oder multispektrale Verfahren mit anschließender Analyse und Bearbeitung der Aufnahmen) oder die Schrift anhand der infolge des früheren Tintenauftrags veränderten Eigenschaften des Pergaments sichtbar gemacht werden kann (insbesondere UV-Fluoreszenzaufnahmen). Voraussetzung ist in ersterem Fall das Vorhandensein von Resten geeigneter Tintenbestandteile im Oberflächenbereich des

² Zu letzteren geben Deckers/Grusková eine Übersicht mit grundlegenden technischen Daten und Berücksichtigung praktischer Aspekte der Anwendung.

Beschreibstoffs, in letzterem die entsprechende Eigenschaft des Pergaments (Fluoreszenz in unbeschriebenen Bereichen). Nicht selten beschränkt sich die Lesbarkeit auf bestimmte Bereiche (z.B. außerhalb des neuen Schriftspiegels) oder auch auf eine der beiden Seiten des Pergamentblatts; oft sind nur einzelne Buchstaben auszumachen. Bei allen Fortschritten der digitalen Aufnahmeverfahren können diese daher meist nur partielle Ergebnisse liefern und erlauben zudem keine Aussage zu den tatsächlich physisch noch vorhandenen Schriftspuren.

Bei den für den Haupttext zum Einsatz kommenden Tinten in dieser Art von Palimpsesten handelte es sich im Regelfall um sogenannte Eisengallustinten. Mit Röntgenfluoreszenzverfahren lassen sich bereits geringe Reste einzelner chemischer Elemente nachweisen; für diese Tinten scheint zunächst insbesondere der Nachweis des Eisens interessant. Dass eine zweidimensionale, punktweise Abtastung eines Palimpsestblatts mit anschließender Darstellung (etwa als Graustufenbild) der erfassten Messwerte insbesondere zum Eisengehalt (sogenanntes »element mapping«) die Wiedergabe auch der Reste der unteren Schrift erlaubt, hat bereits U. Bergmann am Stanford Linear Accelerator bei der Messung einzelner Blätter des Archimedes-Palimpsests gezeigt. Für die Hamburger Untersuchungen entschieden wir uns, jeweils aufgrund punktueller Einzelmessungen bei den eigentlichen Durchläufen für das Mapping eine größere Zahl interessanter Elemente zu berücksichtigen. Als besonders relevant erwiesen sich Eisen, Kupfer und Zink (Eisengallustinte), Calcium (vgl. Auswertung) und Blei (rote Tinten für Initialen, Schmuckelemente usw.).

Während die grundsätzliche Zusammensetzung und das Alterungsverhalten von Eisengallustinten z.B. von Kolar gut untersucht sind (siehe *Iron gall inks*), gibt es insbesondere zu den frühen Tinten, zur detaillierten Entwicklung der Pergamentherstellung oder gar zu den Löschverfahren kaum Erkenntnisse, die über Einzelfälle hinausgehen. Sofern an einzelnen Handschriften insbesondere im 19. Jh. Chemikalien für Versuche zum Einsatz gebracht wurden, getilgte Texte wieder lesbar zu machen, beeinflussen die Rückstände dieser Chemikalien auch den heutigen Befund. Daher sind die zu untersuchenden Palimpseste, genauer gesagt jede einzelne Palimpsesteinheit in einer Palimpsesthandschrift (oft fanden für einen einzelnen neuen Kodex Blätter aus mehr als einer gelöschten Handschrift Verwendung), für unsere Untersuchung jeweils als Unikat zu behandeln. Dieser Ansatz wurde durch die unten vorgestellten Ergebnisse bestätigt.

3. Versuchsaufbau

Für die bisherigen Experimente wurde ein monochromatisierter Photonenstrahl mit einer Anregungsenergie von 18 keV eingesetzt, die Messungen erfolgten mit einem Silizium-Drift-Detektor. Der in Abb. 1 schematisch dargestellte Versuchsaufbau wurde innerhalb der entsprechend abgeschirmten Experimentierstation aufgebaut, in die der

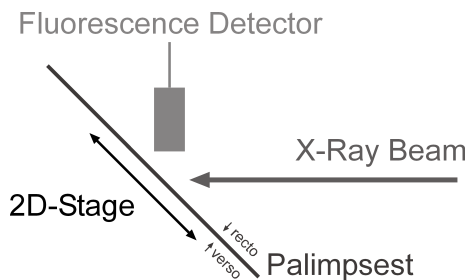


Abbildung 1. Versuchsaufbau (schematisch).

Strahl des DORIS-Speicherrings am HASYLAB über die Hartröntgen-Mikrofluoreszenz-Beamline L geleitet wurde. Das jeweilige Untersuchungsobjekt wurde mit einer Auflösung von ca. 70 Mikrometer in vertikaler und 100 Mikrometer in horizontaler Richtung abgetastet, wobei eine durchschnittliche Abtastgeschwindigkeit von sieben Messpunkten pro Sekunde erzielt werden konnte. Um zunächst eine Vorauswahl der Elemente zu treffen, deren Detektor-Zählraten aus den Sensordaten rechnerisch ermittelt und visualisiert werden sollten, wurden vor den eigentlichen, das ganze Blatt abdeckenden Messungen kleinflächigere Stichproben durchgeführt. Aus den genannten Daten folgt, dass mit den vorliegenden Parametern in einer knapp einwöchigen Messzeit nur der Textbereich von ein bis zwei Blättern (abhängig von ihrer Größe) vollständig abgebildet werden kann.

Da eine gezielte Ablenkung des Röntgenstrahls ab dem Austrittspunkt nur sehr eingeschränkt möglich ist³, war es für die Abtastung erforderlich, die jeweilige Handschrift in einer geeigneten Halterung relativ zum Aufbau aus Strahl und Detektor zu verschieben. Hierzu wurden auf einer Bühne mit ausreichend leistungsstarken Schrittmotoren für die jeweilige Handschrift eigens gefertigte bzw. adaptierte Halterungen eingesetzt, die es ermöglichten, die Handschrift im aufgeschlagenen Zustand zu lagern und das zu untersuchende Blatt so zu fixieren, dass eine Abtastung ohne Beeinträchtigung durch die übrigen Teile der Handschrift möglich war. Auf Grund der Beschaffenheit der getesteten Handschrift musste jedoch der jeweilige Messbereich eingeschränkt werden, so dass z.B. eine Messung bis in Falznähe ausgeschlossen war.

Während die Belastung des Pergaments mit Röntgenstrahlung im für die Untersuchungen erforderlichen Umfang konservatorisch vertretbar ist (Young) und auch geeignete Klimabedingungen in der Messkammer geschaffen und aufrecht erhalten werden konnten, stellte die Konstruktion einer geeigneten Halterung eine Herausforderung

³ Durch den Einsatz spezieller Röntgenspiegel mit einer entsprechenden Steuerung könnte es in Zukunft denkbar sein, dennoch eine der Achsen der Abtastung durch Verschiebung des Strahls abzudecken; dies würde aber eine synchronisierte Verschiebung des Detektors erfordern.

dar. Für einen sicheren Einsatz war ein auch während der Bewegungen durch die Schrittmotoren formstabiler, ausreichend tragfähiger Aufbau zu wählen. Hierzu wurde eine Grundkonstruktion aus Aluminiumprofilblechen gefertigt, die an den erforderlichen Stellen aufgepolstert und mit säurefreiem Papier abgedeckt wurde. Das jeweils zu untersuchende Blatt wurde in einem Rahmen gehalten, in dem es mit Folien stabilisiert werden konnte; dafür kam neutrale Kapton-Folie zum Einsatz. Auf eine Glättung des Blattes wurde verzichtet; durch die Stabilisierung hielt sich eine etwaige Wellung des Blattes allerdings in Grenzen, welche die Messung nicht beeinträchtigten.

4. Auswertung

Ein Blatt des Cod. Lips. gr. 2, das als erstes untersucht wurde, war im 19. Jahrhundert mit sogenanntem roten Blutlaugensalz (Kaliumhexacyanidoferrat(III)) behandelt worden, wovon die Verfärbung des Blattes und Pergamentschäden vor allem im Randbereich zeugen. Bei seiner Untersuchung konnten speziell in solchen Bereichen, die im Zuge früherer Restaurierungsmaßnahmen überklebt worden waren, in der Eisen-Map zusätzliche Teile von Buchstaben dargestellt werden. Zugleich war für viele Teilbereiche ein Fehlen von Schriftresten festzustellen. Das Blutlaugensalz war offensichtlich zeilenweise aufgetragen worden, wodurch insbesondere Eisenspuren im gesamten Auftragsbereich verteilt wurden. Auf diese erste Untersuchung soll hier nicht weiter eingegangen werden.

Detailliertere Ergebnisse werden im Folgenden anhand des Cod. Lips. Rep. I 62 präsentiert. Diese Handschrift enthält zahlreiche Palimpsestblätter, die bei der Untersuchung ohne Hilfsmittel nicht eindeutig früheren Handschrifteneinheiten zugeordnet werden konnten. Eine frühere Behandlung mit Chemikalien ist nicht feststellbar. Unter UV-Licht waren aufgrund der Fluoreszenz vereinzelt wenige Buchstaben lesbar, anhand deren der Text in mehreren Blättern als Johannes Klimakos' *Scala Paradisi* identifiziert werden konnte. Für eines dieser Blätter, auf dem sich zudem eine mit bloßem Auge nicht mehr lesbare Notiz am unteren Rand befindet, wurden mit dem beschriebenen Verfahren Maps von 10 Elementen erstellt⁴. In der bisherigen Teilauswertung zeigte sich, dass das Eisensignal in allen Textbereichen weniger kantenscharfe, stärker verwaschene Ergebnisse als die Kupfer- und Zinksignale lieferte. Die Randnotiz ist in der Kupfer-Map einwandfrei lesbar, der untere Text (dessen Tinte offenbar kaum Spuren von Kupfer enthielt) scheint in der Zink-Map deutlicher erkennbar zu sein, eine endgültige Auswertung steht hier noch aus.

⁴ Nach Abschluss der derzeitigen Versuchsreihe sollen die Visualisierungen der einzelnen Maps bzw. später auch von kombinierten Auswertungen über die Online-Plattform des Teuchos-Zentrums zugänglich gemacht werden.

Da das Röntgenfluoreszenzverfahren die Texte auf beiden Seiten eines Blattes überlagert erfasst, stellt sich die Frage, wie die Texte unterschieden werden können. Auch bei fotografischen Verfahren sind obere und untere Schrift überlagert; eine Trennung in den tatsächlich überlappenden Teilen von Buchstaben ist kaum möglich. Die multispektrale Fotografie bietet hier je nach Tintenbeschaffenheit Unterscheidungsmöglichkeiten, bisherige Resultate lassen allerdings vermuten, dass die überwiegende Zahl der griechischen Palimpseste, in denen untere und obere Schrift mit Eisengallustinten ausgeführt sind, die Voraussetzungen hierfür nicht erfüllen⁵. Beim Element-Mapping zeigte sich bisher in allen Fällen, dass sich die Intensität der Signale zwischen den einzelnen Tintenschichten, und zwar sowohl zwischen oberer und unterer Tinte als auch zwischen denselben Tinten auf den beiden Seiten eines Blattes merklich unterschied. In einem Fall waren diese Intensitätsunterschiede in den visualisierten Daten ohne weitere Bearbeitung der Bilddatei bereits mit bloßem Auge erkennbar. Ob für die überwiegende Zahl der Fälle eine eindeutige Abgrenzung möglich sein wird, lässt sich noch nicht mit Sicherheit sagen, solange weder Vergleichsmessungen in ausreichendem Umfang vorliegen noch weitergehende Bildanalyseverfahren erprobt wurden. Da sich die Texte selten vollflächig überdecken, können allerdings selbst die überlagerten Schriften oft noch entziffert werden.

Der unterschiedliche Befund zur selben Schriftebene auf den beiden Seiten eines Blattes dürfte auf die unterschiedliche Beschaffenheit der Haar- und Fleischseiten des Pergaments zurückzuführen sein, die Unterschiede bei der Haftung und Aufnahme der Tinten bedingt. Für die unteren Texte werden die entsprechenden Unterschiede auch die Wirksamkeit der Löschverfahren beeinflusst haben. Auch aus der optischen Untersuchung von Palimpsesthandschriften war bekannt, dass oft der Text auf einer Seite besser als auf der anderen erkennbar ist. Im seltenen Einzelfall unterschied sich auch die UV-Fluoreszenz beider Seiten so stark, dass bereits daran deutlich unterschiedliche Oberflächenbeschaffenheiten evident waren.

Für das zweite untersuchte Blatt dieser Handschrift (vgl. Abbildungen 2–5) war der untere Text mit optischen Verfahren zuvor nicht identifizierbar. Interessanterweise lieferten hier weder die Eisen- noch die Kupfer-Map deutliche Spuren der unteren Schrift; in der Zink-Map ist eine starke Überlagerung der Tinten feststellbar. Auf diesem Blatt ist die untere Schrift in der Calcium-Map deutlich erkennbar, ein von den bisherigen Untersuchungen abweichender Befund. Es steht zu vermuten, dass im Zuge des ersten Beschriftungsvorgangs verbliebene Spuren des bei der Pergamentherstellung eingesetzten Kalks in die Pergamentoberfläche eingebracht

⁵ Diese und weitere im Artikel genannte Erfahrungswerte gehen auf umfangreiche Arbeiten an Palimpsesten im von Hamburg aus koordinierten EU-Projekt „Rinascimento virtuale – Digitale Palimpsestforschung“ zurück, an denen einer der Verfasser in nicht wenigen Fällen beteiligt war; umfassende quantitative Untersuchungen liegen dagegen nicht vor. Zum Projekt vgl. auch Deckers/Grusková sowie *Rinascimento Virtuale*.



Abbildung 2. Cod. Lips. Rep. I 62, f. 17: Digitalfotografie (recto-Seite).

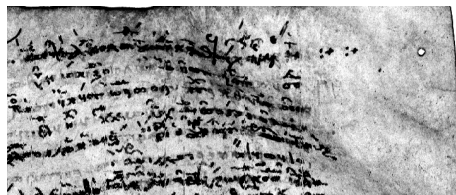


Abbildung 3. Cod. Lips. Rep. I 62, f. 17: Eisen-Map.

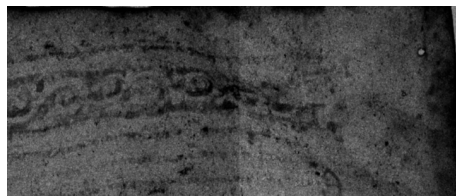


Abbildung 4. Cod. Lips. Rep. I 62, f. 17: Blei-Map.

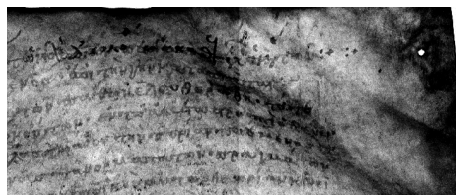


Abbildung 5. Cod. Lips. Rep. I 62, f. 17: Calcium-Map.

wurden und den Löschvorgang überstanden. Es finden sich in der Map nur Spuren einer einseitigen ersten Beschriftung. Da es sich um einen Teil eines fortlaufenden Textes (er konnte jetzt als Teil der *Capita de caritate* des Maximus Confessor identifiziert werden) handelt, ist allerdings als wahrscheinlich anzusehen, dass vor der Löschung eine beidseitige Beschriftung vorlag. Interessant ist, dass die untere Schrift auch im Bereich der roten Zierleiste gleichermaßen lesbar ist; bei optischen Verfahren stellen solche Bereiche oft ein besonderes Problem dar.

Diese Versuche haben gezeigt, dass das beschriebene Verfahren es ermöglicht, Schrift lesbar zu machen, die mit herkömmlichen Ansätzen nicht erkennbar ist. Es kann in vielen Fällen die gelöschte Schrift für wissenschaftliche Zwecke ausreichend darstellen und zugleich den Nachweis ermöglichen, in welchen Bereichen keine Spuren unterer Schrift vorhanden sind. Nachteile sind derzeit der ausschließlich stationäre Einsatz am Ort eines Elektronen-Speicherrings, die erforderliche Untersuchungszeit pro Blatt bei einer nur knapp ausreichenden Ortsauflösung, die Notwendigkeit, jeweils individuelle Lösungen für eine konservatorischen Anforderungen genügende Handschriftenhalterung zu finden sowie ein umständliches und zeitaufwendiges Auswertungsverfahren.

5. Perspektiven

Bei Beginn der hier beschriebenen, über mehrere Jahre erfolgten Untersuchungen war keine mobile Strahlungsquelle verfügbar, die akzeptable Messzeiten ermöglicht hätten. Tests mit neuen Entwicklungen sind für Anfang 2011 vorgesehen. Alternativ wäre auch die eigene Entwicklung eines geeigneten Aufbaus aus mobiler Quelle und verbesserten Detektoren denkbar, die einen mobileren Einsatz ermöglichen würden. Dass dabei freilich nur von einem transportablen, nicht von einem tragbaren Aufbau die Rede sein kann, versteht sich aufgrund der technischen Voraussetzungen von selbst.

Der Zeitfaktor der Messungen ist das zweite entscheidende Problem des jetzigen Aufbaus. Die Flussrate des derzeit genutzten Speicherrings würde nach einer Systemoptimierung eine Erhöhung der pro Sekunde ansteuerbaren Messpunkte um etwa den Faktor 10 erlauben. Die bisherigen Parameter der Messung basieren auf theoretischen Überlegungen und den aus früheren Einzelmessungen vorliegenden Daten. In einer nächsten Messzeit im August 2010 soll anhand umfangreicherer Vergleichsmessungen an bereits untersuchten Proben durch Variation der Energie und der Detektortypen eine Optimierung des Aufbaus erreicht werden. Der nächste Schritt wäre die Entwicklung eines speziell an die Messerfordernisse angepassten, großflächigeren Detektors, um die Signalstärke weiter zu erhöhen.

Eine weitere deutliche Verkürzung der Messzeiten in Hamburg wird auch durch einen Wechsel vom bisher eingesetzten Speicherring DORIS zum neuentwickelten Ring

PETRA erreicht werden können, der eine erhebliche höhere Flussrate liefert. Diese kann freilich nur in Verbindung mit einem geeigneten Detektor und entsprechend schneller Ausleseelektronik nutzbar gemacht werden. Mit einer solchen Erhöhung der Geschwindigkeit wäre auch eine Erhöhung der Ortsauflösung möglich; je nach Handschrift wäre pro Richtung eine Verdoppelung wünschenswert.

Wenn aufgrund entsprechender Entwicklungen in den kommenden Jahren die Untersuchung mehrerer Seiten an einem Tag ermöglicht werden kann, wäre das vollständige Mapping ganzer Palimpsesthandschriften in akzeptablen Zeiträumen durchführbar, für das dann auch die Einrichtung eines festen Messaufbaus wünschenswert wäre. Ein solches Szenario hätte den Vorteil, dass im Zusammenhang einer vollständigen Untersuchung gemeinsam mit den handschriftenbesitzenden Bibliotheken auch über erforderliche konservatorische bzw. restauratorische Maßnahmen für die jeweiligen Handschriften nachgedacht werden könnte und unter optimalen Bedingungen eine vollständige Untersuchung (z.B. auch einschließlich einer herkömmlichen Digitalisierung usw.) der Einzelblätter vorgenommen werden könnte. Sobald absehbar ist, dass mit einem dauerhafteren Messaufbau eine größere Zahl von Handschriften untersucht werden kann, wäre auch die Konstruktion darauf abgestimmter universellerer Halterungen in zwei oder drei Grundgrößen denkbar, die an die jeweiligen Handschriften oder Einzelblätter anpassbar sind.

Auch im Bereich der Auswertungsverfahren sind Weiterentwicklungen in Arbeit, einschließlich der Trennung der verschiedenen Schriftebenen anhand abweichender Signalstärken und der Kombination der verschiedenen Element-Maps (z.B. mittels sogenannter blind source separation). Die an jedem Messpunkt aufgenommenen Röntgenfluoreszenzspektren liefern zudem die Daten für eine Elementanalyse der verwendeten Tinten, aus denen sich als Nebenprodukt Durchschnittswerte für die gesamte Seite bilden lassen.

Bibliographie

Bergmann, Uwe. »X-Ray Fluorescence Imaging of the Archimedes Palimpsest: A Technical Summary«. Stanford (CA): Stanford University, 2006.

<http://www.slac.stanford.edu/gen/com/images/technical%20summary_final.pdf>.

Deckers, Daniel und Jana Grusková. »Zum Einsatz verschiedener digitaler Verfahren in der Palimpsestforschung« *The legacy of Bernard de Montfaucon: Three Hundred Years of Studies on Greek Handwriting. Proceedings of the Seventh International Colloquium of Greek Palaeography (Madrid – Salamanca, 15–20 September 2008)*. Eds. Antonio Bravo García und Inmaculada Pérez Martín, with the assistance of Juan Signes Codoner. Bibliologia 31. Turnhout: Brepols, 2010 353–362 [im Druck].

Iron gall inks: on manufacture, characterisation, degradation and stabilisation. Hg. Kolar, Jana und Matija Strlič. Ljubljana: Narodna in univerzitetna knjižnica, 2006.

- Rinascimento Virtuale. Digitale Palimpsestforschung. Rediscovering written records of a hidden European cultural heritage.* A project for the rediscovery and dissemination of Greek palimpsests. Universität Hamburg, 2001–2010. <<http://www.rinascimentovirtuale.eu/>>.
- Young, Gregory. »Effect of High Flux X-radiation on Parchment«. Report No. Proteus 92195. Ottawa: Canadian Conservation Institute, 2005.
<http://www.archimedespalimpsest.org/pdf/archimedes_f.pdf>.

Counting Sheep: Potential Applications of DNA Analysis to the Study of Medieval Parchment Production

Timothy Stinson

Abstract

This chapter follows up on several preliminary tests that have shown that DNA survives in medieval parchment manuscript leaves and may be extracted and analyzed, and offers suggestions for defining and implementing future genetic studies of parchment. It articulates the need to consider genetic data in conjunction with other types of evidence—such as historical texts and archaeological data—both in planning tests of parchment and in interpreting the results of such tests. I consider the potential influences of diet, urbanization, market and trade specialization, and changes in agricultural practices and animal husbandry on parchment production, and discuss how genetic analysis can contribute to our knowledge of these topics as well as how historical and archaeological evidence will both complicate and contextualize data derived from genetic testing.

Zusammenfassung

Dieser Beitrag schließt sich an eine Reihe von ersten Tests an, die gezeigt haben, dass DNA in mittelalterlichen pergamentenen Handschriftenseiten nachweisbar ist und extrahiert und analysiert werden kann. Es werden daraus Anregungen für die Bestimmung und Umsetzung zukünftiger genetischer Pergamentuntersuchungen entwickelt. Genetische Daten, so wird argumentiert, werden dabei im Zusammenhang mit anderen Erkenntnissen, etwa aus historischen Texten oder archäologischen Befunden, sowohl bei der Planung kodikologischer Untersuchungen als auch bei der Interpretation ihrer Ergebnisse eine wichtige Rolle spielen. Der Beitrag diskutiert die möglichen Einflüsse von Nahrung, Urbanisierung, der Spezialisierung des Handels sowie von Wandlungen in bäuerlichen Praktiken und der Haustierhaltung auf die Pergamentproduktion. Besprochen wird auch, wie genetische Analysen zum Erkenntnisgewinn in diesen Fragen beitragen und ob und wie sie historische und archäologische Belege sowohl in Frage stellen als auch kontextualisieren.

1. Introduction

In December of 2009, I published an article in *The Papers of the Bibliographical Society of America* (PBSA) detailing the successful outcome of a project designed to determine the feasibility of extracting the DNA contained in the parchment leaves of medieval manuscripts (Stinson). Working with C. Michael Stinson, a biologist with experience in phylogenetic research, I was able to confirm that DNA not only survives in medieval parchment, but also that it may be extracted and analyzed in order to reveal bibliographical information as well as information about the animals whose skins furnished the parchment. Through an analysis of the DNA contained in five disbound parchment leaves that had likely once belonged to a fifteenth-century Flemish book of hours, we were able to determine the genetic relatedness of calves that had been used in the manufacture of the book, showing that two leaves were derived from one maternal lineage and the other three from a second maternal lineage. Our project joins a small number of similar studies that seek to leverage the genetic information contained in parchment in order to gain a fuller understanding of medieval books and documents in their historical and physical contexts; Woodward et al. conducted a study of the parchment of the Dead Sea Scrolls, Burger et al. surveyed a wide variety of pre-historic hided materials, and Poulakakis et al. published a study of the parchment of Greek manuscripts written on goat skin.¹ All of these studies have been exploratory in nature, and each has demonstrated the survival of DNA in medieval parchment and the feasibility of extracting it in order to identify the species from which the parchment was made and to reveal the genetic relatedness (or lack thereof) of the animals used to make it. Our study is ongoing, with current work focusing on the development of techniques for testing parchment that are minimally destructive but provide reliable scientific results.²

In addition to detailing the techniques and results of our study, my article briefly suggested four potential benefits that the analysis of genetic information offers to the study of medieval books, which I repeat here in abbreviated form:

1. Localizing herds. Once a substantial number of manuscripts were tested and the results entered into databases, we would have the potential to localize both herds

¹ See also Spencer and Howe (246–47), for a brief discussion of their experimentations “with methods for extracting DNA from parchment”. To date there have been no further publications from their study.

² All studies thus far have used pieces of parchment or hide material larger than is permissible or desirable for the large-scale study of library and archival collections of manuscripts. Our own study used samples measuring 0.5×0.5 cm taken from the center of the lower margin of manuscript leaves, a sample size that is ample from scientific terms, but clearly not ideal from a preservation or curation point of view. A further goal of the current phase of our project is the analysis of nuclear as well as mitochondrial DNA, which will be necessary in order to achieve some of the possibilities mentioned in this essay. The distinction between these types of DNA is discussed in my PBSA article.

and manuscripts. Working from manuscripts with known dates and provenance,³ we might be able to construct models showing the likely family descent and local origin of animals and parchment, thereby equipping scholars with a new tool for determining the origins of manuscripts.

2. Studying the parchment trade. Little is known about the medieval parchment trade, a deficit that this project could begin to remedy. It is generally assumed that early in the medieval era, monasteries that were engaged both in copying texts and in raising herd animals for meat, hide, and wool likely used their own herd animals to produce parchment, and that the earlier a manuscript is the more likely that this is the case. Commercial production of parchment in the later medieval era, meanwhile, likely involved trade routes and the mixing and redistribution of skins in both market towns and in workshops producing and selling parchment in large cities and university towns such as London, Paris, and Oxford. Such practices would complicate the study of both herd populations and the origins of individual skins, but conversely parchment from this era might provide data for tracking the movement of animals and the trade and distribution of skins.

3. Analyzing the construction of codices. Although parchment is quite durable, as are the books created from it, many manuscripts have of course deteriorated substantially due to poor handling, intentional disbinding for financial or other motives, fire, moisture, overzealous binding, and misguided conservation efforts. This has often resulted in the loss of codicological information necessary to understand how the manuscript was initially constructed. Because bifolia were originally one piece of skin, DNA identification to the level of the individual organism might allow scholars to deduce the original gatherings and how they were combined to create a codex.

4. Resolving debates concerning individual manuscripts. Scores of puzzles and debates surrounding single codices might be resolved (or at least one position in these debates substantiated) through such analysis. A good example of such, and one that has been previously raised by Christopher de Hamel ("DNA – Genetic Fingerprinting"), is the famous Bury Bible, one of the treasures of the Parker Library at Corpus Christi College, Cambridge that dates to the twelfth century. The *Gesta Sacristarum*, a late thirteenth-century history of the Bury St. Edmunds abbey where the Bury Bible was made, states that the parchment used for the Bible's illustrations was a special, expensive lot brought in "from regions of *Scotia*" because Master Hugo, the illuminator, could find no local parchment to suit him (*Memorials* ii, 290). This story is seemingly borne out by the fact that the illuminations are all rendered on individual leaves of parchment glued to ones beneath them that are sewn into the book, but scholars disagree on whether *Scotia* refers to modern-day Scotland or Ireland. An analysis of the origins of this

³ A good starting point for such research would be standard reference works listing dated and datable manuscripts such as Robinson's *Catalogue of Dated and Datable Manuscripts c. 888–1600 in London Libraries* and Watson's *Catalogue of Dated and Datable Manuscripts c. 435–1600 in Oxford Libraries*.

book combined with comparisons to parchment from other codices could augment our understanding of a medieval historical text (the *Gesta*), add important context to our understanding of how a medieval Latin place name was used, and go far towards solving a modern scholarly debate about this important manuscript.

The third and fourth benefits listed above are reasonably self-evident; if we are able to identify the animals from which parchment leaves were derived as well as their potential relatedness to other animals in the same book, then it is clear that this information might provide direct information about how a disbound book was initially constructed and/or to contextualize the construction of a given book (e.g., all animals in the book are closely related and from a definable area, or the book is an admixture of more distantly related animals). My goal in this essay is to discuss the first two potential benefits in greater detail. I will consider how genetic studies of parchment might make the realization of these goals possible as well as how we will need to consider genetic data in conjunction with other types of evidence—such as historical texts and archaeological data—both in planning tests of parchment and in interpreting the results of such tests. Without doing so, we will be unable to reach the fullest possible understanding of the significance and meaning of genetic data found in medieval parchment.

2. Parchment as Archaeological Evidence

It is worth pausing first to consider several ways in which parchment provides particularly unusual and valuable evidence concerning human use of and interaction with animals. To date, parchment's historical value has primarily been understood to be its role as a substrate for written texts; its durability and ability to support and preserve texts and illustrations over many centuries have contributed enormously to the survival and transmission of much of our intellectual and cultural heritage from the medieval era. Viewed another way, we might see parchment leaves themselves as important artifacts that contain much information about human industry as well as the lives of the animals from which the substance was derived. It is not unusual for animal remains from the medieval era to be examined for evidence of what they can tell us about medieval life; for example, zooarchaeologists exhume bones, shells, and teeth in order to understand how animals were utilized, processed, and disposed of by humans, while museum curators and historians study decorative arts and practical everyday objects made from the same materials, from musical instruments to combs and knife handles. But parchment, which survives in tens of thousands of extant medieval books, many of which contains hundreds of pieces of parchment each, as well as in countless unbound documents held in libraries and repositories around the world, remains for the most part unexamined as archaeological evidence. There is, of course, a good reason for this, the very reason that so much parchment has survived—much of it in

excellent condition—while many other artifacts from the Middle Ages survive only in poor condition, if at all. This is the fact that the books containing the parchment are highly prized both for their aesthetic merits and for the information conveyed in the texts they comprise, and many have been considered worthy of curation and careful treatment from early in their histories (although many others have not, unfortunately). As Holsinger notes, “[w]hile a phylogenetic archaeologist will not hesitate to extract a nice chunk of ossified bone marrow from a thirteenth-century cow femur for laboratory analysis, the book is priceless; the book [...] has legal protection” (620).

We are fortunate indeed that books have been so prized, for animal bone and hide were used to manufacture many common items in the medieval era, including clothing, footwear, bags, drinking vessels, and bone tools of many varieties, but these items range from being relatively rare to virtually impossible to find today. Parchment books survive in very large numbers, however, and the development of techniques for extracting and analyzing DNA means that the parchment leaves that these books contain now have many stories to tell other than those written upon them, and the survival of so many samples in good condition offers a wealth of possibilities for scientists to explore once sufficiently non-invasive testing procedures have been developed.⁴ Medieval parchment books are especially important sites of archaeological and historical evidence because of their nonpareil combination of physical and textual information. A cow femur dug from an archaeological site will not have written upon it the name of someone who processed or consumed the cow, nor will a comb made from bone usually have written upon it the date and place of its origin, but similar information is routinely found in certain types of texts such as legal documents and chronicles. We cannot assume that parchment was derived from the same area where texts were copied onto it, as discussed below, but the combination of local historical information such as names, dates, place names, and regional dialects with the very large-scale survival of well-preserved artifacts containing DNA is without parallel, and we stand to gain much from analyzing this information in tandem.

DNA evidence is particularly promising due to its potential to be combined with other types of evidence, including not only historical and legal texts, but also other faunal remains, to develop much more specific and complete understandings of how medieval humans interacted with and used animals during the medieval era. In order to understand how such studies of genetic information might usefully be conceived and carried out, it is first necessary to acknowledge the ever-shifting nature of the conditions of humans and domestic animals across the medieval era. Population growth,

⁴ Developing such techniques is the focus of the next phase of our ongoing research into the genetic analysis of parchment. Possible techniques include taking very small core samples from parchment leaves and swabbing the surface of the parchment or rubbing it gently with a metal implement in order to get useable cells. Techniques for taking tiny core samples have not yet been perfected, and swabbing or rubbing the surface increases the likelihood of contaminated samples.

changes in economic structures, shifts in the types of food consumed, climate change, deforestation, increasing numbers of roads and opportunities for trade, new agricultural techniques and practices, and the hunting of animals such as bears and wolves to near extinction all had profound impacts on humans as well as the animals on which they depended, whether wild or domestic (Resl Introduction 3). Furthermore, such changes happened at different times in different areas of Europe, and were subject to other local conditions, including regional climate, natural disasters, war, and disease. The production of parchment for books was impacted by all of these forces (and the demand for parchment may have in turn influenced at least some of them, such as agricultural practices and trade), and as such we should be hesitant to accept blanket assertions about its production, value, and use across the whole of medieval Europe.

3. Genetics and the Parchment Trade

The two topics that I have identified for this study—the genetics of local herds and the parchment trade—are necessarily intertwined; in order to trace how parchment moved from place to place, we will first need to be able to identify local and regional groups of animals, which seems feasible due to the fact that in Europe parchment was almost always produced from domesticated herd animals. As a case study, I will consider here one question regarding parchment that has been frequently posed—how expensive was it to produce a single book?—and focus primarily on one animal and nation—the sheep in England—although I will draw upon evidence of other animals and countries for purposes of comparison. This will demonstrate not only how genetic information might be used to answer such a question, but also how other disciplines and types of evidence that inform us of herd populations and the parchment trade will be necessary to contextualize genetic information. My aim here is to think through some of the preliminary conditions for designing a study of medieval parchment and book production that incorporates genetic evidence, including what other disciplines are already able to tell us regarding these topics at any given time and place in history, for without doing so we are unlikely to arrive at meaningful results simply by taking decontextualized samples of genetic materials from medieval books.

In order to answer the question of how expensive it was to produce a book copied on sheepskin parchment in medieval England, we must consider how many skins such a book would require as well the relative value of those skins if used in other ways (e.g., leather) or the value of the animal if left alive (e.g., wool, milk). As Febvre and Martin noted some decades ago, a number of written accounts of the number of animal skins used in the production of parchment books provide exaggerated claims that do not square with basic mathematical calculations:

A simple calculation will be sufficient to demolish the stories so often told about the fabulous numbers of sheep and calves required to make a single book. Even modern works of scholarship continue to repeat these old errors. Thompson, for example, quotes an order by the Countess of Clare in England in 1324 for a copy of the *Vitae Patrum*, for which no fewer than 1,000 skins were allegedly required. At the current price of 2 pennies per skin, the vellum alone would have cost the fabulous sum of £6. In fact an examination of the *Vitae Patrum*, whether in Latin or in one of the contemporary French versions, quickly shows that when written in two columns the text generally fills about 150–160 leaves of 25 cm × 16 cm, a total area amounting to no more than 6 square metres—a dozen skins at the most. (17)⁵

Even if such clearly inaccurate calculations are too common, they may at least easily be shown to be in error. A more complicated matter is the commonly repeated and potentially misleading claim that a single book might comprise the skins of an entire herd or flock; for example, Jean Leclercq notes that “a flock of sheep was needed to provide the parchment necessary for copying a book by Seneca or Cicero” (123). Such claims tend to obscure the truth as much as they relate it, for they simply convey into the reader’s mind whatever size they imagine a medieval flock of sheep to be, whereas the reality varied from a few sheep owned by a private farmer to herds comprising tens of thousands of animals owned by noblemen or industrious monastic orders. For example, “[t]he Benedictines of Ely were already feeding 13,000 sheep at the time of the Domesday Book (1086),” and the Benedictines at Winchester Cathedral Priory are recorded as owning a flock of 20,000 sheep in 1320 (although of course these may have been dispersed into numerous smaller flocks), which could provide enough parchment for many books year after year (Butler and Given-Wilson 85–6). This is not to say that very large volumes did not utilize the skins of many animals; the largest volumes utilized one calfskin per bifolium, as these were of course larger than the skins of young sheep. Bruce-Mitford calculated that the Codex Amiatinus—which measures 505 × 780 mm, has more than 1,000 folios, and weighs approximately 75 pounds—contains 515 calfskins (2). The Book of Kells, meanwhile, measures 330 × 240 mm and contains 340 folios and was made from as many as 150 calves, and even so it is incomplete; approximately 40 folios are missing and a binder has trimmed the leaves down from an estimated original size of 370 × 260 mm (Henry 152). Thus claims that one flock (or herd) of animals went into one book must be considered carefully, as a small flock might number a dozen sheep—and a small book be produced from that dozen—while even the largest books do not approach the sizes of the largest flocks of many thousands of domesticated animals.

⁵ Febvre and Martin here cite J. W. Thompson, *The Medieval Library* (645).

Since books and herds vary greatly in size in medieval England, it may be more useful to consider the value of a single animal skin at any given time, and to extrapolate from that the value of the parchment in one book. One way to do this, of course, is to consult records for sales of unfinished skins and/or finished parchment, as Febvre and Martin do in the passage quoted above; they argue that in both England and France the price of parchment remained “reasonably stable from the second half of the 14th to the first half of the 15th century, when book production was increasing rapidly, and this seems to prove that it was not such a rare commodity” (18). The authors are here interested in the impact of printing, and the subsequent increase in the number of books being created, on the price of parchment, and researchers should follow their lead in consulting records of the price and sale of parchment for the time periods and local regions that interest them and match their research agenda. But such records are not always available, and determining the relative value of a skin in their absence is a complex matter, for its value must be considered in relation to other uses for either it or the animal that had to be killed in order to provide it:

Looked at another way, what was the economic cost to the community to produce (i) the raw material, and (ii) the secondary product? Was the cost very high in relation to alternative uses for the skins, or in terms of the decision to slaughter an animal before it had reached full maturity and therefore maximum meat weight? Or did parchment-making, with its demand for young animal skins, fit easily into a system where many young male animals were slaughtered annually in order to conserve winter fodder for the breeding and milk-yielding females? (Ryan 125)

As Ryan notes, determining the relative value of a skin involves imagining two other possible scenarios. The first is that the raw skin is put to alternative uses, such as the production of leather. The second is the economic benefit of letting an animal reach adulthood. Because the skins of young animals not old enough to reproduce were used for parchment production, such benefits may be numerous, including the ability of these animals to reproduce (thus providing offspring with their own economic benefits), the increased meat that might be provided by adult animals, larger hides for leather (although these would be too thick for parchment), wool, manure for fertilizer, horn (which was used for a variety of manufactured goods), and milk. These economic benefits of letting animals reach an adult age were potentially offset by other factors, including a market for veal or lamb, the economic benefit of having fresh meat to sell throughout the year, sufficient grazing space, the fact that adult hides were more likely to be damaged by disease or insects, and the cost of providing fodder for adult animals during the winter in colder climates.

Genetic analysis has the potential to answer clearly and definitively the number of different animals whose skins were used to produce the parchment in a given book,

such as our hypothetical English book with sheepskin leaves, for tests can be conducted in such a way that they distinguish one organism from another. It is possible, of course, that one skin would have been made into parchment sheets that ended up in different volumes—indeed it must have happened frequently—but calculations that estimate the surface area of one skin combined with genetic evidence of how many different animals are found in one volume should provide a very close estimate of the number of individual animals used in any given volume, as well as how likely it is that part of these skins were used in other volumes or for other purposes. Determining the relative value of these skins, however, as well as the economic tradeoffs that would have been made in culling young animals for parchment instead of letting them reach maturity, cannot be achieved from genetic analysis alone, and it is here that combining research into DNA with extant scholarship will be crucial.

4. Uses of Historical and Archaeological Evidence

In his introduction to *Breaking and Shaping Beastly Bodies: Animals as Material Culture in the Middle Ages*, an anthology of essays on zooarchaeological topics, Terry O'Connor notes that archaeologists "need all of the other disciplines that study the medieval world" in order to understand and contextualize their data (6). Similarly, in the course of reporting his study of domestic animal remains from the medieval era found near Dudley Castle, Richard Thomas notes that "[z]ooarchaeologists have tended to use historical facts as 'interesting anecdotes' rather than as an integral part of research", while "the majority of historical expositions regarding the exploitation of animals in the medieval and post-medieval periods are considered almost to the complete exclusion of archaeological evidence" (17–18). In order to realize the full potential of genetic analysis of parchment, and to avoid producing reports of isolated facts about the genetic makeup of medieval sheep, cattle, and goats, we need to heed such advice by considering the extant archaeological and historical data both before and after conducting genetic tests of parchment. I would thus like to turn now to examples of the types of historical and archaeological evidence that might help to answer questions regarding the relative value of parchment (as well as many other questions), evidence that genetic analysis of parchment might serve to contextualize and that in turn might serve to clarify and make meaningful the results found through DNA testing.

A particularly important body of such evidence lies in our knowledge of medieval practices of animal husbandry. I have already noted that herd sizes differed significantly in the medieval era, which would of course have implications for the genetic variety of any given herd population. As is still the case today, many domesticated male animals were castrated, which made them more docile and their meat more tender. A few adult males were of course needed for breeding, but then as now one bull or ram would be

sufficient for a large numbers of cows or ewes. As the size of the herd increases, we should of course expect to find more genetic diversity, as a small herd might descend from one adult male for a number of consecutive years, whereas a large herd numbering thousands of animals would be more likely to result in a breeding female mating with different males in consecutive years, and would necessitate more than one breeding male for the herd each year. In addition to varieties in the size of herds, there are general trends in animal husbandry across the time period that also must have affected the genetic composition of those herds. In early medieval England, for example, sheep and other domesticated animals were “semiwild” and lived outdoors “between the farm and forest”, requiring protection from forest predators, but “[t]he general trend between the eleventh and fifteenth centuries brought animals from the forest to the farm and from the farm to the urban market and slaughterhouse” (Pascua 82–3). This process culminated at the end of the Middle Ages with the policy of enclosure, which maximized the income of wealthy landowners, but famously led to depopulation of villages, unemployment, and other social ills, as well as to a sheep population that “had multiplied so rapidly that it produced a major crisis in the use of land” (Lander 38). The story of sheep populations in England is not one of uninterrupted expansion, however, as changes in climate and outbreaks of disease periodically decimated populations:

The best-known early epidemics occurred during the period 1315–1319. However, the problems began in the decade of 1270–1280. From England to Castile, chroniclers mention declining crop yields and monasteries, particularly Cistercian, unable to produce the amount of wool contracted with merchants. Large manors went bankrupt, a sign of an economy in collapse. The flocks of the Bishop of Winchester, which exceeded 27,000 sheep in 1272, numbered fewer than 9,000 in 1278 and yet fewer in 1280. (Pascua 96)

Selective breeding was also practiced and developed during the Middle Ages in England; the Cistercians, for example, were “pioneers in grading their wools, which the Benedictines had preferred to sell mixed and in bulk” and “studied feeding and breeding, and the possibility of grappling with the deadly disease of sheep-rot” (Butler and Given-Wilson 85). The potential of genetic studies of parchment to contribute to our knowledge of animal husbandry practices is clear. It may be possible, for example, to articulate genetic differences between free-roaming sheep that were kept by shepherds—and perhaps mated with other herds found “between the farm and forest”—and those sheep kept in enclosures at the end of the medieval era. Or perhaps we may be able one day to chart the influence of Cistercian practices on breeds of sheep. But this information also serves as a caution to us, for it shows that we cannot assume that something true of ninth-century animals will also be true of late fifteenth-century animals, and thus we must proceed carefully both in designing our genetic studies of parchment and in drawing conclusions from those studies. The variety of situations outlined here also

shows how the relative value of parchment might have shifted over time. The amount of labor required to shepherd sheep versus keeping them in enclosures, the waxing or waning market demand for wool over the centuries, and the overpopulation (or underpopulation) of sheep might all impact the relative value of letting lambs reach adulthood versus the value of culling them for parchment.

Another factor that likely had an effect on the value of parchment is dietary practice and the relative abundance or scarcity of food, which, as with other things we have seen, differed throughout time and from place to place. The impact of culling young animals from species also used for meat and dairy is obvious: they provide less meat than if they had grown to their full weight, and females culled at a young age do not produce milk and do not have offspring that may also be used for meat and dairy products. The direct impact of diet on parchment production, or vice versa, however, is difficult and perhaps impossible to discern, barring the discovery of parchment and bones from butchered animals that share an ancestry.⁶ But since skinning animals and butchering them necessarily go hand in hand, and since the respective needs for meat and skins must have had impact on one another, it is worth reviewing relevant historical and archaeological evidence of using herd animals as meat sources in medieval Europe in order to acquire a fuller understanding of how culling these animals for skins must have been integrated into annual agricultural cycles and practices that also produced sufficient food for medieval people. I will consider three relevant topics: who ate meat and what sort they ate, annual cycles of butchering and preserving meat, and regional varieties in such practices.

As with many things in medieval society, access to meat was strongly affected by social class and status:

Food was also class specific in the Middle Ages. The ability to access and afford foodstuffs and clothing materials of different types was generally regulated by economic constraints. The diet of peasants continued largely to be based on cereals and tended to feature meat only if they could hunt it down, whereas

⁶ In certain research situations, such a discovery may be less fortuitous than it sounds. Woodward and his colleagues working on the Dead Sea Scrolls were able to analyze samples from ibex and goat hides and also “to isolate and amplify DNA from archaeological bones of ibex and goats found at Masada”, thereby demonstrating their ability “to recover the necessary genetic information from ancient animal remains that will enable [...] comparisons between the scroll fragments and the animals from which they were derived” (Woodward 228). Many excavations of the remains of butchered domestic herd animals in Europe involve very large numbers of bones; see, e.g., Maltby and Buglione, whose excavations each yielded data from thousands of bones and/or horn cores. Buglione, meanwhile, differentiates between the bones of mature animals and those killed at twelve months or younger in Apulia, a distinction that could be very useful to genetic studies of parchment. She notes that in late antiquity 3.5% of cattle were killed under 12 months of age, and in the early Middle Ages the number was 17.30%, whereas with sheep the numbers are 38.8% and 32.7% respectively (194–95).

the aristocracy consumed meat on a regular basis, with special treats reserved for feast days. (Resl Introduction 5)

It should be remembered that peasants would not have been owners of large herds of domestic animals, and thus that decisions regarding how much meat a herd should produce would likely reflect the dietary needs and habits of a smaller elite minority than that of an entire local population; not only veal and lamb, but also beef and mutton would have been rare treats for many peasants throughout medieval England. Members of monastic orders that raised herd animals would have had their own supply of meat, but they had dietary regulations that, if followed, would have reduced their consumption of this meat. In a study of monks in late medieval Westminster, for example, Barbara Harvey notes that their diets were subject to special restrictions during periods of fast, and that “[o]utside the fast season of Advent and Lent, an average week in the monastery comprised four meat days and three fish days, and in principle every monk ate flesh-meat on two of the meats days and meaty dishes on the other two” (63). Moreover some monastic orders abstained from eating meat because it “was clearly identified as being the penchant of a certain echelon of society” from which they wished to distance themselves (Seetah 25, Bond 77).

Records from monastic houses and manorial kitchens sometimes offer information useful in determining the types and quantities of meat consumed. Harvey documents a wide variety of fish and meat sources used at the monastery in Westminster, including thirteen types of fish, chickens, ducks, geese, conies, and mature and young sheep, cattle, and pigs (tables A and B, 226–28). Historical information concerning who had access to meat as a food source and how much those individuals consumed helps to contextualize parchment production because it shows that only a limited portion of the population had rights to or created demand for meat from these animals, that this population had access to a wide variety of other animal food sources, and, in the case of monastic orders, that consumption of meat was itself limited due to the periods of fasting tied to the liturgical calendar as well as weekly dietary guidelines (and of course some of these fasts and dietary guidelines would have applied to observant laymen as well). Records kept by managers of manorial and monastic kitchens also permit scholars to estimate the number and size of animals butchered for food, thereby providing further context for understanding how the use of animals for meat might have related to the parchment trade. For example, Harvey provides estimates that a mature cow carcass weighed 308 pounds, whereas a calf weighed fifty-seven pounds; an adult sheep, meanwhile, weighed thirty-one pounds whereas a lamb weighed eleven and a half. Such evidence not only suggests how large—and thus how old—young sheep and cattle were when culled, perhaps for parchment, but also provides a basis for calculating how large their skins would have been.

Medieval consumption of meat also varied seasonally, as the only alternatives to fresh meat were salted and smoked meats. Pork, which cured better than beef or mutton, was subject to a “seasonal chronology of butchery”, with chins (“the backbone and immediately adjoining area”) being consumed as early as October through December, preserved meat being eaten through the spring, and fresh meat available thereafter (Woolgar 116–17). A similar chronology is found in the consumption of other mammals. Throughout much of medieval Europe, a large slaughter of animals in the autumn, often on Martinmas (November 11), not only provided a source for salted meat, but reduced the number of animals that would need to consume fodder over the course of the winter:

The Martinmas slaughter of animals, salting down carcasses, continued as a method of ensuring a supply of preserved meat through the winter. It was practised at Frampton in 1343, at Hunstanton in 1349 and in many other places. Throughout the period, fresh meat was available in winter, but it was more expensive and its consumption was restricted to those of the highest status. By the mid-fifteenth century, a higher proportion of cattle may have been available as a consequence of the driving trade, bringing cattle from the north and west—and the availability of fresh meat increased. (Woolgar 112–13)

The large-scale slaughter of sheep and cattle at Martinmas suggests the expense, and perhaps impossibility, of keeping many adult herd animals alive over the winter, especially in colder climates. This, in turn, suggests that the culling of animals earlier in the year—at a time when their skins would be suitable for producing parchment—would have fit into this annual cycle in a way that would not necessarily mean that culling these young animals was an enormous financial sacrifice. For many of the animals, the only options seem to have been culling them for fresh meat and skin suitable for parchment production or killing them only a few months later and salting the meat (and indeed some of the animals slaughtered at Martinmas would still have been young enough for their skins to be suitable for parchment production).

It should be noted that dietary preferences, the availability of meat as a food source (whether fresh or preserved), and both the ability to preserve food as well as the pressures to do so due to impending winter weather varied from place to place and from time to another across medieval Europe. For example, Pascua notes both an overall increase in meat consumption in Europe among the lower classes during the fourteenth and fifteenth centuries and regional differences in what meats were available and/or preferred:

The diet of thirteenth-century peasants consisted mainly of bread and dairy produce, that is, cheese and milk, which together accounted for four-fifths of the calorific value of all food consumed. Fowl was the main source of meat

in France, and pork was the main source of meat in Britain and Germany. In the fourteenth and fifteenth centuries, animal protein accounted for 40 percent of total food. The common diet was based on mutton and goat in the Mediterranean region, Italy, and Spain; fresh beef in Hungary, the Low Countries, and Sweden, and pork in France and Germany. Beef and mutton remained paramount in Britain, where ovine livestock predominated, though less so than in the south of Europe. (98)

The reasons for these changes in diet have implications for studies of parchment, for Pascua notes that they occurred because the “key developments set in motion by the economic growth of the central centuries of the Middle Ages persisted: urbanization, market integration, and regional specialization” (98). Urban markets demanded that agricultural products, including meats and hides, be brought in from the countryside, and large towns and cities had regions that specialized in working animal products. The first stop for animals in specialized market systems would have been the butcher.⁷ After butchering, parts of the animals were frequently distributed to a number of specialist workers, including furriers, tanners, parchmenters, pinners, glovemakers, and others, many of which were organized into guilds. As such, evidence for the local origins of these animal parts may be obscured due to these manufacturing activities, which would have mixed together and redistributed very large numbers of animal carcasses. Transhumance, meanwhile, meant that many animals traveled significant distances before they were slaughtered. For example, cattle drovers regularly brought cattle from Scotland and Wales into England (Pascua 98).

In addition to the transhumance of livestock to meet market demands for meat, there was also significant trade of finished animal products along established trading networks throughout medieval Europe. According to Veale, “[t]he anonymous author of the political poem, *The Libelle of Englyshe Polycye*, writing between 1436 and 1438, commented on Ireland’s great wealth in skins, referring to the good martens, deer, otter, squirrel, hare, sheep, lamb, fox, kid, and rabbit skins with which she traded”, while Scotland, meanwhile, had a “flourishing trade in fox, squirrel, marten, cat, beaver, and otter skins” (60). England was (and still is) famous for its wool, which found a ready market not only in market towns within England, but in continental centers specializing in cloth production, such as those in Flanders and Florence (Butler and Given-Wilson 85). Transhumance was also practiced for the purposes of maintaining flocks of sheep kept for their wool; in Spain, over three million Merino sheep were involved in an “unceasing flow from north to south” along “complex networks of routes that peppered the landscapes” of the Iberian Peninsula (Pascua 94). It is likely that parchment moved along some of these same trade routes for furs and wool. A very promising possibility is that genetic analysis of skins might help to trace the movement of parchment from

⁷ See Seetah for an overview of the importance and development of the butcher’s trade in the medieval era.

one region to another, but this evidence of robust international trade must also give us pause, since it suggests that we cannot simply assume that parchment dated through textual evidence to a particular locality was manufactured there, or that the animals from which it was made originated there.

Finally, it is worth considering that the broad trends that I have outlined here concerning sizes of herds and flocks, animal husbandry, diet, the movement of animals and goods manufactured from them, urbanization and the development of market economies, and regional differences in diet and the animals preferred as either livestock or food were all subject to disruption at any time from forces such as war, epidemic disease of either humans or livestock, drought, flood, and other factors. For example, Ryan reports that salt, which was not locally available in Ireland, was periodically scarce: “In A.D. 1338, a rise in the price of salt was recorded in Clyn’s Annals, and in 1486, the chronicler of the Annals of Ulster recorded a severe shortage” (137). This, of course, would have implications for salting beef to preserve it over the winter, and might result in more animals being eaten at a younger age and their hides made available for the production of parchment or leather. Bad weather in “1315–1317 laid the foundation for endemic murrains that affected bovines and sheep from Ireland to Germany”, and a few years later “[r]eference to catastrophic animal mortality and ruined crops in every monastic cartulary suggests rates of mortality in oxen of between 25 and 50 percent and between 50 and 70 percent in sheep” (Pascua 96–7). And of course wars and catastrophic losses from plague affected the human populations of much of Europe during this century as well. Any of these events could have drastic and perhaps long-lasting impact on the value of animals and their skins, and on the relative merits of culling animals early enough to produce parchment.

5. Conclusion

The experiment that I described in the *PBSA* article and the experiments by Woodward, Burger, and Poulakakis have shown conclusively that DNA survives in medieval parchment and that it may be extracted and analyzed. We are faced now with the many millions of extant surviving parchment pages and how we might approach unlocking their secrets through genetic analysis. Such broad-scale analysis is predicated upon developing minimally destructive techniques, but this is a matter of when and not if; one only needs to consider the miracle that the possibilities of this technology would appear to be to medieval parchment makers—or even nineteenth-century parchment makers—to have confidence that the development of such techniques will come with time. The question, then, will become how best to deploy the technology, and what sorts of questions we might pose and answer with it. I have attempted to show here both the significant possibilities and the potential complications of genetic analysis of

parchment. On the one hand, we have the potential to leverage the genetic information of parchment to obtain unparalleled glimpses into the medieval past. Perhaps we will be able to trace parchment trade routes or document the effects of the enclosure system, or Cistercian breeding agenda, or salt shortages in Ireland, or epidemic murrains across Europe on the production of books and the lives of humans and animals from many centuries ago. But the conditions that I have outlined here also serve as a caution that we will not be able to discern the full meaning of the genetic information contained in parchment if we do not consider it alongside historical and archaeological information (much of which itself remains insufficiently examined). Isolated genetic information from single leaves, or even single books, will in most cases remain unclear, and may be misleading, without a historical understanding of trade and agricultural practices; although genetic data will likely greatly enrich such fields of study, it will also rely upon them, especially in the early stages. In designing studies of the genetic data contained in parchment and in interpreting the results of them, we must always be mindful of the ever-shifting landscape of parchment production in the medieval world.

Bibliography

- Bond, James. "Production and Consumption of Food and Drink in the Medieval Monastery." *Monastic Archaeology: Papers on the Study of Medieval Monasteries*. Oxford: Oxbow, 2001. 54–87.
- Bruce-Mitford, Rupert L. S. *The Art of the Codex Amiatinus*. Jarrow: Parish of Jarrow, St. Paul's House, 1978. [Jarrow Lecture, 1967.]
- Buglione, Antonella. "People and Animals in Northern Apulia from Late Antiquity to the Early Middle Ages: Some Considerations." Pluskowski 189–216.
- Burger, Joachim, et al. "Mitochondrial and Nuclear DNA from (Pre)historic Hide-derived Material." *Ancient Biomolecules* 3 (2001): 227–38.
- Butler, Lionel and Chris Given-Wilson. *Medieval Monasteries of Great Britain*. London: Michael Joseph, 1979.
- "DNA – Genetic Fingerprinting of Medieval Manuscripts," *University of Cambridge, Corpus Christi Alumni News*. September 2003.
- Febvre, Lucien and Henri-Jean Martin. *The Coming of the Book: The Impact of Printing, 1450–1800*. Trans. David Gerard. London: Verso, 1990.
- Harvey, Barbara. *Living and Dying in England, 1100–1540: The Monastic Experience*. Oxford: Clarendon Press, 1993.
- Henry, Françoise. *The Book of Kells: Reproductions from the Manuscript in Trinity College, Dublin*. New York: Alfred A. Knopf, 1974.
- Holsinger, Bruce. "Of Pigs and Parchment: Medieval Studies and the Coming of the Animal." *Publications of the Modern Language Association of America*. 124.2 (2009): 616–23.
- Lander, Jack R. *Conflict and Stability in Fifteenth-Century England*. London: Hutchinson University Library, 1969.

- Maltby, Mark. "Urban-Rural Variations in the Butchering of Cattle in Romano-British Hampshire." *Diet and Craft in Towns: The Evidence of Animal Remains from the Roman to the Post-Medieval Periods*. Ed. D. Serjeantson and T. Waldron. Oxford: B.A.R., 1989. 75–106. BAR British series 199.
- Memorials of St. Edmund's Abbey*. Ed. Thomas Arnold. London: Eyre and Spottiswoode, 1890–96. 3 vols. *Rerum Britannicarum Medii Aevi Scriptores, Rolls Series*, No. 96.
- O'Connor, Terry. "Thinking About Beastly Bodies." Pluskowski 1–10.
- Pascua, Esther. "From Forest to Farm and Town: Domestic Animals from ca. 1000 to ca. 1450." Resl, *Cultural History* 81–102.
- Pluskowski, Aleksander, ed. *Breaking and Shaping Beastly Bodies: Animals as Material Culture in the Middle Ages*. Oxford: Oxbow, 2007.
- Poulakakis, Nikos, et al. "Ancient DNA and the Genetic Signature of Ancient Greek Manuscripts", *Journal of Archaeological Science* 20 (2006): 1–6.
- Resl, Brigitte, ed. *A Cultural History of Animals in the Medieval Age*. Oxford: Berg, 2007. Print. Vol. 2 of *A Cultural History of Animals*. Linda Kalof and Resl, gen. eds. 6 vols. 2007.
- Resl, Brigitte. "Introduction: Animals in Culture, ca. 1000–ca. 1400." Resl, *Cultural History* 1–26.
- Robinson, Pamela R. *Catalogue of Dated and Datable Manuscripts c. 888–1600 in London Libraries*. London: British Library, 2003. 2 vols.
- Ryan, Kathleen. "Parchment as Faunal Record." *MASCA: University of Pennsylvania Journal* 4.3 (1987): 124–38.
- Seetah, Krish. "The Middle Ages on the Block: Animals, Guilds, and Meat in the Medieval Period." Pluskowski 18–31.
- Spencer, Matthew and Christopher J. Howe. "Authenticity of Ancient-DNA Results: A Statistical Approach." *The American Journal of Human Genetics* 75.2 (2004): 240–250.
- Stinson, Timothy L. "Knowledge of the Flesh: Using DNA Analysis to Unlock Bibliographical Secrets of Medieval Parchment". *The Papers of the Bibliographical Society of America* 103.4 (2009): 435–53.
- Thomas, Richard. "Of Books and Bones: The Integration of Historical and Zooarchaeological Evidence in the Study of Medieval Animal Husbandry." *Integrating Zooarchaeology*. Ed. Mark Maltby. Oxford: Oxbow Books, 2006. 17–26.
- Thompson, James Westfall. *The Medieval Library*. New York: Hafner, 1939.
- Veale, Elspeth M. *The English Fur Trade in the Later Middle Ages*. 2nd ed. London: London Record Society, 2003. London Record Society Publications, v. 38.
- Watson, Andrew G. *Catalogue of Dated and Datable Manuscripts c. 435–1600 in Oxford Libraries*. Oxford: Clarendon Press, 1984. 2 vols.
- Woodward, Scott R., et al., "Analysis of Parchment Fragments from the Judean Desert Using DNA Techniques." *Current Research and Technological Developments on the Dead Sea Scrolls*. Ed. Donald W. Parry and Stephen D. Ricks. Leiden: E. J. Brill, 1996. 215–38. *Studies on the Texts of the Desert of Judah*, v. 20.
- Woolgar, Christopher M. *The Great Household in Late Medieval England*. New Haven: Yale UP, 1999.

Thermographie – ein neuartiges Verfahren zur exakten Abnahme, Identifizierung und digitalen Archivierung von Wasserzeichen in mittelalterlichen und frühneuzeitlichen Papierhandschriften, -zeichnungen und -drucken

Peter Meinlschmidt, Carmen Kämmerer, Volker Märgner

Zusammenfassung

Der Beitrag präsentiert die Thermographie als ein neues Forschungsinstrument der Wasserzeichenkunde. Einführend wird zunächst die Bedeutung von Wasserzeichen für die Erschließung historischer Dokumente erörtert. Darauf aufbauend stellen die Autoren ein neuartiges Verfahren zur Visualisierung von Wasserzeichen vor, durch welches es in einfacher Weise gelingt, auch überschriebene oder übertuschte Wasserzeichen klar und deutlich sichtbar zu machen. Durch diese neue Technik ist es möglich, Wasserzeichen von historischen Dokumenten mittels einer Thermographiekamera in einer der Buchdigitalisierung ähnlichen, einfachen Weise und *in situ* abzunehmen. Anschließend werden einige Möglichkeiten vorgestellt, die mit Hilfe der modernen Bildverarbeitung und Mustererkennung die vergleichende Suche nach identischen und ähnlichen Wasserzeichen erleichtern.

Abstract

This chapter introduces thermography as a new research instrument for watermark studies. As an introduction, the importance of watermarks for the analysis of historical documents is discussed in detail. On this basis, a newly developed technique for the visualisation of watermarks is introduced. With this technique, it is relatively easy to identify watermarks even if the structures are obscured by ink or pigments. The newly developed thermographic method allows identification and storage of watermarks *in situ* without damaging the paper in a similarly easy way as today's scanning of books. As an outlook, the authors present how modern image processing and pattern recognition can support the comparative search for identical or similar watermarks.

1. Einleitung

Die Wasserzeichenkunde oder Filigranologie gilt heutzutage innerhalb der Erschließung von historischen Buchbeständen, Karten, Graphiken, Zeichnungen etc. als unverzichtbare historische Hilfswissenschaft (Hay). Doch nicht nur aus papier- und buchhistorischer Perspektive stellen die Wasserzeichen einen zentralen Forschungsgegenstand dar: Sie werden zum Treffpunkt der Disziplinen, wenn es darum geht, Abnahmeverfahren zu optimieren und Motivzuordnungen durch Mustererkennung zu automatisieren. Die Synergieeffekte interdisziplinären Arbeitens auf diesem Gebiet sollen durch diesen Beitrag dokumentiert werden. Die im Folgenden vorgestellten Ansätze verstehen sich als komplementär zur derzeit auch andernorts in Projektarbeit betriebenen Wasserzeichendokumentation und -erschließung (Wolf).

Der Untersuchungsgegenstand sind Wasserzeichen in handgeschöpften europäischen Büttenpapieren, wie sie seit dem 13. Jh. nachzuweisen sind (Weiß). Wasserzeichen definieren sich nicht nur über ihr Motiv, sondern auch durch das Siebgeflecht, auf dem sie befestigt waren und das sich ebenfalls als »Wasserzeichen« im historischen Handpapier erhalten hat. Sowohl Wasserzeichenmotiv als auch Siebgeflecht sind oft im Durchlicht erkennbar. Die Position einer Drahtfigur auf dem Schöpfsieb bestimmt sich in ihrem Verhältnis zu den Ripp- und Kettdrähten, auf denen sie befestigt war¹ (Klinke). Es gibt Wasserzeichen, die zwischen zwei Kettdrähten befestigt waren und andere, bei denen der Kettdraht die Wasserzeichenfigur der Länge nach durchteilt (Piccard 1956 101). So ist auch die exakte Dokumentation des das Wasserzeichenmotiv umgebenden Drahtgeflechts zur Identifizierung von Wasserzeichen unabdingbar. Das angewandte Abnahmeverfahren muss daher das Wasserzeichen in seiner Gesamtheit so exakt wie möglich wiedergeben.

Stellt man die Frage nach der Bedeutung von Wasserzeichen, so ist diese zum einen für die Vergangenheit und zum anderen für die Gegenwart zu beantworten. Wasserzeichen dienten zunächst als Firmenlogos und Qualitätszeichen für verschiedene Papiersorten der einzelnen Papiermühlen. Es verband sich mit den Wasserzeichen aber auch das Bestreben nach Fälschungssicherheit und Qualitätssicherung. Dieser Umstand schlägt sich im »Tractatus de insignis et armis« des Rechtsgelehrten Bartolus de Saxoferrato (1313/14–1357) nieder (Lackner). Bartolus de Saxoferrato will die Verwendung bestimmter Wasserzeichenmotive für den Inhaber einer Papiermühle rechtlich geregelt wissen. Andere Papiermüller durften gleiche Motive nicht verwenden. Wenn man also die Wasserzeichen eines Papiermüllers identifizieren kann, lassen sich bestimmte Papiere regional zuordnen und Handelswege sowie die Verbreitung von Papier nachvollziehen. Beispielsweise verwendete die Papiermühle in Bern das Wahrzeichen der Stadt, den Bären, als Wasserzeichenmotiv. Jaffé verweist im

¹ Die Ripp- oder Bodendrähte verlaufen rechtwinklig zur Schmalseite des Schöpfsiebes. Die Kett- oder Bindedrähte verlaufen parallel zur Schmalseite des Schöpfsiebes.

Zusammenhang auf die drei Qualitätsstufen für eine Ravensburger Papiermühle. Papier erster Güte wurde dort in der Mitte des 15. Jhs. mit dem Bild eines Tores mit zwei Türmen gekennzeichnet. Die zweite und dritte Qualitätsstufe markierte der Ochsenkopf ohne Augen mit einem darüber stehenden Kreuzstab und ein Hifthorn. Neben der Kennzeichnung der Papierqualität dienten die Wasserzeichen auch der Markierung verschiedener Papierformate bzw. Bogengrößen.

Die Wahl des Motives verband sich oftmals auch mit einer besonderen Assoziation. So ist das Ochsenkopfwasserzeichen in großer Vielfalt nachgewiesen (Jaffé 23). Es handelt sich hierbei um ein Handwerkersignet, da es als eine Referenz an den Evangelisten Lukas zu interpretieren ist, dessen Tierattribut, der Ochse, den Papiermachern als Schutzheiliger galt. Die gesamte mittelalterliche Bilderwelt und Symbolik spiegelt sich in der unglaublichen Vielfalt der Wasserzeichenmotive wider und bedarf unter verschiedenerlei Hinsicht noch einer genaueren Erforschung.

In der Gegenwart besitzen die Wasserzeichen unter unterschiedlichen Aspekten zentrale Bedeutung. Zum einen dienen sie der Datierung bisher undatierten Schriftguts. Die Wasserzeichenrepertorien von Charles Moïse Briquet (1839–1918) und Gerhard Piccard (1909–1989) sowie die im »Bernstein-Portal« (Bernstein-Projekt) zusammengeführten Wasserzeichendatenbanken bieten unter dieser Fragestellung z.B. der Handschriftenerschließung ein zentrales Instrumentarium. Piccard (1965) kommt aufgrund von sehr differenzierten Vergleichen und Reihenaufstellungen zu dem Ergebnis, dass sich der Beschriftungszeitraum für Papiere aus der Zeit von 1370 bis 1630 auf nicht mehr als drei oder vier Jahre beläuft. Entsprechend können dann undatierte Papiere aufgrund von in den Repertorien und Datenbanken nachgewiesenen Wasserzeichenbelegen datiert werden. Auf diese Weise gelingt außerdem die Beweisführung für Alters- und Echtheitsbestimmungen von Kunstwerken. Aber auch die Bestimmung der zeitlichen Reihenfolge, in der Johann Sebastian Bach seine Werke komponiert hat, ließe sich durch die Wasserzeichenidentifikation und dadurch ermöglichte Papierdatierungen der betreffenden Musikhandschriften bewerkstelligen (Weiß). Für die Katalogisierung und Erschließung von mittelalterlichen Handschriften geben die Wasserzeichen Anhaltspunkte zur Formatbestimmung: Vollständige Wasserzeichen verweisen auf Folioformate, durch die Bindung einmal geteilte Wasserzeichen auf Quartoformate usw.

Da bisher nur von der Beweisführung anhand identischer Wasserzeichen die Rede war, muss nun auch auf Wasserzeichenvarianten eingegangen werden. Die Erkenntnis über die Bedeutung von Wasserzeichenvarianten hat die Forschung moderner Abnahmeverfahren wie beispielsweise der Betaradiographie² (Kushel 1999) und der Thermographie (Neuheuser) zu verdanken, an die aufgrund ihrer Exaktheit und ihrer Möglichkeit, vielfältigeres und umfänglicheres Datenmaterial miteinander zu

² Betaradiographie ist eine Technik bei der die durch radioaktiven Zerfall entstehenden Betastrahlen zur Durchstrahlung von Papier genutzt werden. Gemessen werden mit dieser Methode die Dicke, der Wassergehalt, die Dichte aber auch als bildgebendes Verfahren die Wasserzeichen.

vergleichen, die Durchzeichnungsmethoden Piccards und anderer Wasserzeichenexperten nicht heranreichen³. Bedient man sich fotografischer Abnahmeverfahren, die das genaue Umfeld des Wasserzeichenmotivs in einer größeren Detailvielfalt wiedergeben, können Varianten eines Wasserzeichens, die z.B. durch Abnutzung der Drahtfigur entstanden sind, durchaus zum Vergleich und somit zur zeitlichen Einordnung von Schriftstücken etc. herangezogen werden (Haidinger 13; Hedges). Alle Kett- und Rippdrähte werden abgebildet, so dass die Aussage über Identität oder Variabilität eines Wasserzeichens differenzierter wird. Beispielsweise können unterschiedliche Papierdicken und entsprechend unterschiedlich starke Transparenz bei einzelnen Ripp- oder Kettdrähten als Anhaltspunkte betrachtet werden. Auch leichte Verschiebungen der Drähte sind durch fotografische Verfahren exakt wiedergebbar. Diese Verfahren ermöglichen außerdem die Erstellung von chronologischen Reihen, die sich aus dem unterschiedlichen Abnutzungsgrad der Drahtfiguren ergeben. Auf diese Weise werden Varianten erkennbar, die dann zur Datierung herangezogen werden können. In einem zweiten Schritt, nämlich durch die Mustererkennung, lassen sich die Parameter von Wasserzeichendefekten, Siebdefekten, Siebpreparaturen, Wasserzeichenverschiebungen u.Ä. definieren. Die Belastungen des Schöpfsiebes beim Abgautschen des geschöpften Papierbogens und auch später beim täglichen Reinigen des Siebes mit Bürsten bewirkten diese Deformationen und machten in einigen Fällen auch eine Reparatur notwendig. Ausbrüche und Fehlstellen in der Drahtfigur sind in Wasserzeichensammlungen, die mit fotografischen Verfahren erstellt wurden, eindeutig nachweisbar und gelten als wichtige Indizien in der Wasserzeichenforschung.

Im Folgenden wird nun auf das Abnahmeverfahren der Thermographie eingegangen, da diesem besonders unter dem konservatorischen Gesichtspunkt Bedeutung zugemessen wird⁴. Die Thermographie eignet sich, wie oben erwähnt, außerordentlich gut zur Dokumentation von Wasserzeichenvarianten. Sie bringt den großen Vorteil mit sich, dass bei stark beschriebenen Papieren und bestimmter Tintenzusammensetzung die Schrift in der Aufnahme ausgeblendet und so das Wasserzeichen vollständig erkennbar wird. Im Vergleich zu radiographischen Abnahmeverfahren ist die Thermographie ungleich kostengünstiger, mit weniger gesundheitlichen Risiken für die Bearbeiter und weniger Zeitaufwand bei der Erstellung einer Aufnahme verbunden. Ein großer Vorteil beim Anlegen von Wasserzeichensammlungen mittels Thermographie beruht auf der Tatsache, dass die Aufnahmen gleich in digitaler Form vorliegen und so leicht bearbeitbar und archivierbar sind.

Die Vorteile des nachstehend beschriebenen Verfahrens zur Mustererkennung liegen in der o.g. Parametrisierung von Einzelcharakteristika der Wasserzeichenbelege, die

³ Eine Übersicht zu den verschiedenen Abnahmeverfahren von Wasserzeichen ist bei Kämmerer (43–47) zu finden.

⁴ Andere Verfahren, wie z.B. die Durchreibung werden als konservatorisch bedenklich eingestuft, da es sich nicht um ein berührungsfreies Abnahmeverfahren handelt.



Abbildung 1. Foto von einer Rembrandt-Handzeichnung im Auflicht (links) und im Durchlicht (rechts) (Döhring 117, 183). Aus der rechten Abbildung wird ersichtlich, dass Wasserzeichen auch im Durchlicht oft kaum erkennbar sind.

dann einen automatisierten Vergleich ermöglichen. Dies hat zur Folge, dass eine exakte Beschreibung von Einzelbelegen unabhängig von einer mitunter trotz terminologischer Standardisierungsleistung uneindeutigen sprachlichen Beschreibung möglich wird. Dieser Aspekt ist von besonderer Relevanz bei komplexen Wasserzeichen, die sich aus mehreren Elementen zusammensetzen. Beispielsweise spielt die Reihenfolge der Einzelelementbezeichnungen bei der automatischen Mustererkennung keine Rolle. Auf diese Weise trägt das unten beschriebene Verfahren zusätzlich zu einer exakten Wasserzeichenerschließung bei.

2. Visualisierung von Wasserzeichen

Die Visualisierung von Wasserzeichen erscheint auf den ersten Blick sehr einfach. Man hält das Papier zwischen eine Lichtquelle und das Auge und kann mehr oder weniger gut das Wasserzeichen erkennen. Fehlende oder nicht sichtbare Bereiche werden durch das subjektive Wissen des Bearbeiters um und die Erinnerung an ähnliche Wasserzeichen oder durch die Imagination ersetzt. Doch bereits das Fotografieren dieses Wasserzeichens führt dazu, dass einige Bereiche durch Dickenunterschiede des Papiers oder durch Schriftzeichen oder Übermalungen nicht so kontrastreich aufgenommen werden können wie bei modernen, sehr homogenen Papieren ohne störende Beschriftung (Abbildung 1 und 2).

Durch die Entwicklung neuer digitaler Kameras und Bildverarbeitungstechniken wurde die Digitalisierung historischer Texte und Bücher und deren Wasserzeichen stark beschleunigt. So wurden bereits sehr früh verschiedene Kamertypen (Abbildung 2) und Filter genutzt, um unerwünschte Überlagerungen zu reduzieren.



Abbildung 2. Foto eines unsignierten, beidseitig beschriebenen historischen Dokuments von 1811⁵ in reflektiertem Streiflicht (links) und im Durchlicht (Mitte). Das rechte Foto zeigt einen vergrößerten Ausschnitt des im Durchlicht unter der Eisengallustinte schwer sichtbaren Wasserzeichens.

Die Fotos in Abbildung 2 wurden mit einer CCD Kamera aufgenommen, deren spektrale Empfindlichkeit vorwiegend im sichtbaren Wellenlängenbereich zwischen 400 nm und 900 nm liegt. Im Gegensatz zu den Fotos in Abbildung 2, die mit breitbandigem, sichtbarem Licht aufgenommen wurden, sind in Abbildung 3 dieselben Fotos nur mit einem schmalbandigen Bandpass-Filter bei 780 nm (links) und bei 830 nm (rechts) aufgenommen worden.

Deutlich sind in den Aufnahmen mit der höheren Wellenlänge weniger Überlagerungen durch die Eisengallustinte zu beobachten und es ist damit ein besserer Kontrast im Bild zu erreichen – das Wasserzeichen ist besser zu erkennen, auch wenn es immer noch von störenden Schriftzeichen überlagert ist.

Diese Beobachtungen legen nahe, Kameras zu nutzen, die im nahen Infrarot-Wellenlängenbereich ihre größte Empfindlichkeit besitzen, um die Überlagerungen z.B. durch Eisengallustinte zu unterdrücken (Kushel 1985). Trotzdem verbleiben oft unerwünschte »Geisterbilder«, die das Wasserzeichen stören.

⁵ Das Dokument stammt aus dem Privatbesitz von Dr. Märgner.

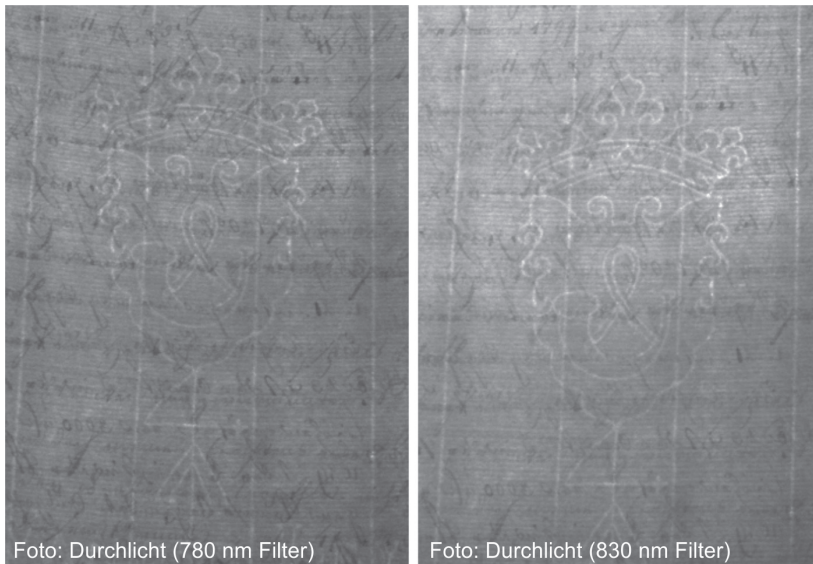


Abbildung 3. Foto des historischen Dokuments aus Abb. 2 im Durchlicht bei der Wellenlänge von 780 nm (links) und 830 nm (rechts).

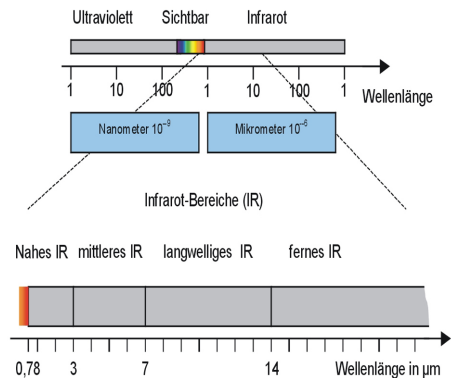


Abbildung 4. Aufteilung des Infrarot-Spektrums in die verschiedenen Wellenlängenbereiche (Meinlschmidt).

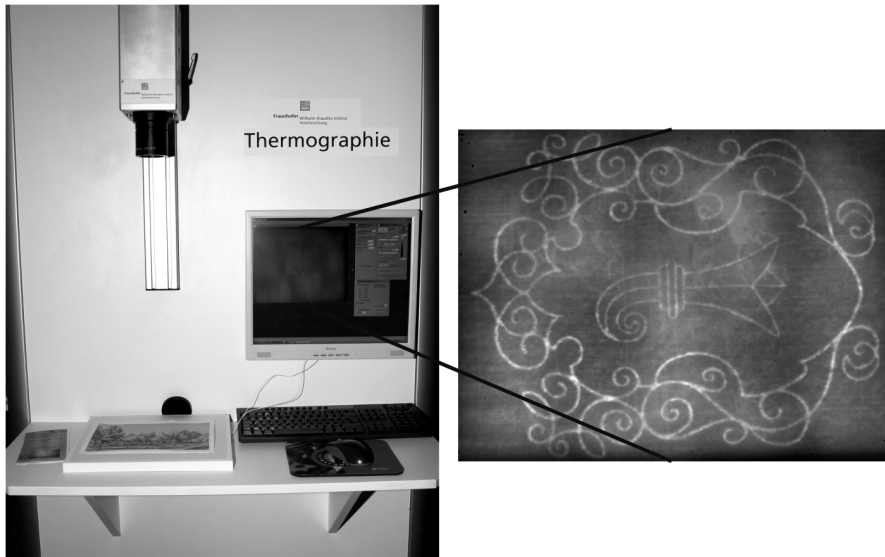


Abbildung 5. Foto der Aufnahmetechnik mit Thermographiekamera, Wärmeplatte und der auf einem Passe-Partout liegenden Handzeichnung (links). Rechts ist das Wasserzeichen der Handzeichnung ohne störende Überlagerungen zu sehen.

3. Visualisierung von Wasserzeichen mit thermischer Infrarot-Strahlung

Im Gegensatz zum nahen Infrarot-Spektrum (Abbildung 4) mit Wellenlängen von $0,7 \mu\text{m}$ bis $3 \mu\text{m}$ reicht das thermische mittlere Infrarot-Spektrum (MWIR) von $3 \mu\text{m}$ bis $7 \mu\text{m}$ und das langwellige Infrarot-Spektrum von $7 \mu\text{m}$ bis $14 \mu\text{m}$ (LWIR).

Die Kameras, die entweder im MWIR- oder im LWIR-Bereich empfindlich sind, werden auch Thermographiekameras oder Wärmebildkameras genannt und können Temperaturunterschiede von wenigen hundertstel Grad Celsius ($<0,02 \text{ }^\circ\text{C}$) wahrnehmen. Die Bilder dieser Kameras (Thermogramme) sind Schwarz-Weiß-Bilder mit einer Informationstiefe von 12 bis 16 Bit Graustufen, die oft zum besseren intuitiven Verständnis in Falschfarben konvertiert werden, bei denen blaue und schwarze Farben kalte und rote Farben warme Bereiche darstellen.

Eine neue Infrarot-Technik zur Visualisierung von Wasserzeichen wurde in Kooperation zwischen dem Fraunhofer-Institut für Holzforschung (WKI) und dem Institut für Nachrichtentechnik (IfN) der Technischen Universität Braunschweig entwickelt (Meinschmidt und Märgner).



Abbildung 6. Aufbau zur thermographischen Aufnahme von Wasserzeichen in einzelnen Seiten aus einem historischen Buch⁶.

Die Methode beruht einerseits darauf, dass viele verschiedene Tinten und Druckfarben im Infrarot-Wellenlängenbereich transparent sind und andererseits diese Strahlung an Papier und Wasserzeichen unterschiedlich absorbiert und gestreut wird.

Für erste Experimente wurde eine Thermographiekamera über einer Wärmeplatte positioniert, die mit einem Passe-Partout als Rahmen versehen war. Auf diesem wurde das Untersuchungsobjekt platziert, ohne die Wärmeplatte zu berühren (Abbildung 5).

Wird die Heizplatte erwärmt, emittiert sie wie ein »Schwarzer Strahler« (Maldague) Infrarot-Strahlung, die das Papier durchdringt und von der Thermographiekamera darüber aufgezeichnet wird.

Da die Strahlung je nach Dicke und Struktur des Papiers unterschiedlich absorbiert bzw. gestreut wird, kann das Wasserzeichen sofort auf einem Monitor beobachtet und auf einem Computer aufgezeichnet werden.

Während der Aufbau in Abbildung 5 für die Untersuchung einzelner Blätter geeignet ist, muss für die Untersuchung eines historischen Buches, insbesondere bei nur sehr kleinen erlaubten Öffnungswinkeln, eine andere Konfiguration verwendet werden (Abbildung 6).

Das Thermogramm in Abbildung 7 (links) zeigt ein Wasserzeichen, das mit der neuen Infrarot-Technik in Abbildung 6 aufgenommen wurde, während es sich in

⁶ Königswinter, Pfarre St. Remigius, Missale, fol 154r, Foto: Fraunhofer WKI, Braunschweig.

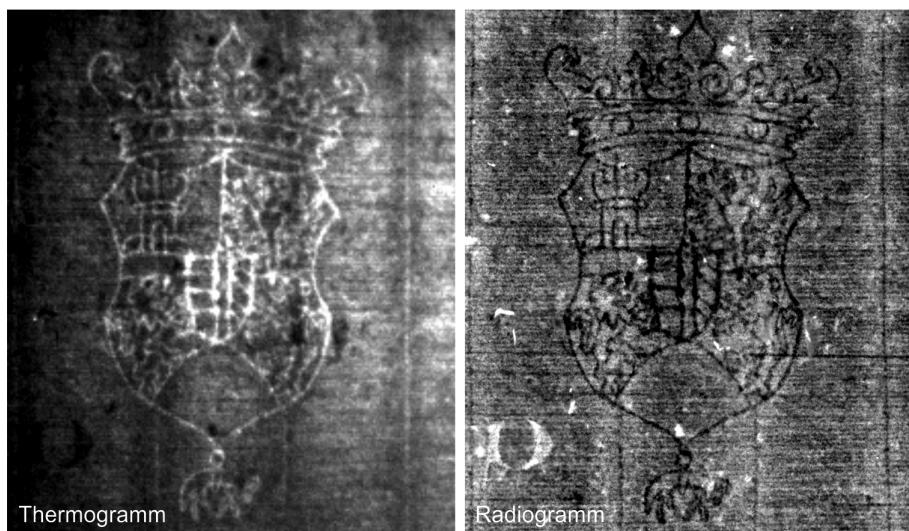


Abbildung 7. Infrarot-Bild (links) und Röntgenbild (rechts) eines Wasserzeichens aus dem Buch⁸ in Abb. 6.

der rechten Abbildung um eine Röntgenaufnahme⁷ desselben Blattes handelt. Der Kontrast der beiden Wasserzeichen ist ähnlich gut, jedoch sind die Grauwerte durch die unterschiedlichen Techniken invertiert (Neuheuser u.a.).

Für diese Messung wurde die Heizplatte in Abbildung 5 durch eine schwarz mattierte, dünne Kupferplatte ersetzt und von der Rückseite innerhalb von Millisekunden mit einer Blitzlampe erwärmt. Während dieser kurzfristigen Erwärmung wurde die durch das Papier transmittierte Infrarot-Strahlung mit einer Bildaufnahmefrequenz von 130 Hz als Videosequenz aufgezeichnet. Das Bild mit dem besten Kontrast ist in Abbildung 7 (links) zu sehen.

Die Thermogramme wurden mit einer Kamera aufgenommen, die 384×288 Pixel besitzt und mit einem Stirlingkühler auf eine Temperatur von -196°C abgekühlt wurde,

⁷ Aufnahme an einer 320kV-Anlage, kleiner Brennfleck; Röhrenspannung: 290 kV; Dosis: 5 mA; Aufnahmerichtung: Sagittalaufnahme anteriorposterior; Film-Fokus-Abstand: 800 mm; Metallfolie: Blei; Filmtyp: D 4 doppelseitig; Belichtungszeit 2 Minuten (mit Variation einer zweiten Bleifolie zur posterioren Abschwächung); resp. Filmtyp: D 3 einschichtig; Belichtungszeit: 2 Minuten (mit Variation eines 10 mm-Kupfer-Vorfilters). Das Ergebnis waren normale Röntgennegative, welche mit einer Auflösung von $50\ \mu\text{m}$ Pixelgröße von einem Laser-Röntgenfilmsscanner digitalisiert wurden.

⁸ Elektronenradiographische Aufnahme eines Wappen-Wasserzeichens, erkennbar ist zudem eine Initiale in Minium (Königswinter). Foto: Berlin, Bundesanstalt für Materialforschung und -prüfung.

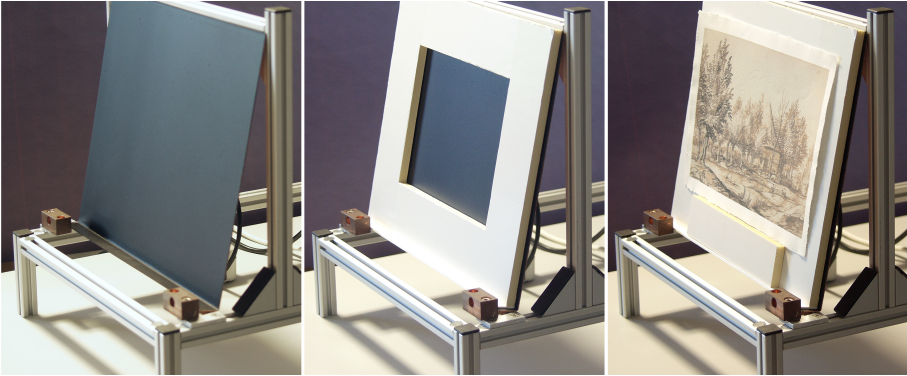


Abbildung 8. Eine schwarz lackierte Kupferplatte mit einer Temperatur von 40 °C dient als Infrarot-Strahler (links), ein Passe-Partout (Mitte) dient zur Einhaltung des passenden Abstands zwischen dem wertvollen Bild und der Heizplatte (rechts).

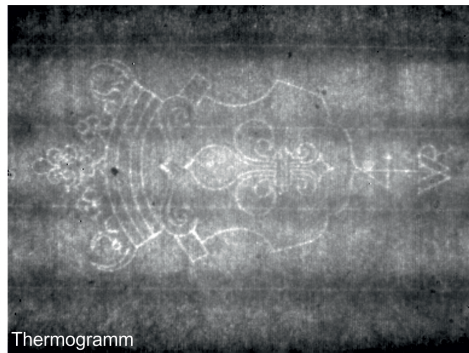


Abbildung 9. Links ein Detail des Bildes aus Abb. 8 (rechts), das zur Kollektion der Alten Meister des Herzog Anton-Ulrich Museums in Braunschweig gehört. Auf der rechten Seite ist das zum Bild gehörende Wasserzeichen zu sehen, das mit der Infrarot-Kamera ohne Interferenz durch die Zeichnung aufgenommen wurde.

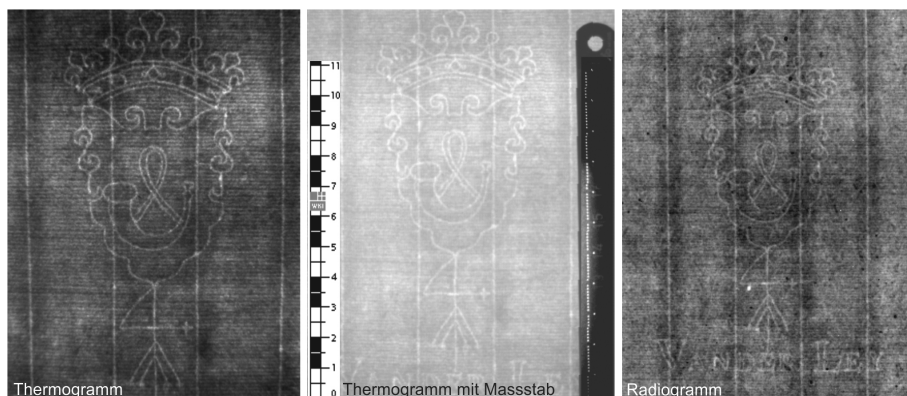


Abbildung 10. Wasserzeichen des Papiers in Abb. 2, aufgenommen mit Infrarot-Technik (links), mit einer Skalierung (Mitte) und mit Röntgentechnik der Firma Viscom (rechts)⁹.

so dass die Ortsauflösung ca. $25\ \mu\text{m}$ bei einer Temperaturlösung von ca. 20 mK (Ircam; Thermosensorik) erreicht.

Bei der mit Blitzlampen erwärmten Kupferplatte wird der Kontrast und damit die Erkennbarkeit des Wasserzeichens im Wesentlichen durch die Inhomogenität der Hintergrundstrahlung bestimmt. Um den Kontrast zu verbessern, wurde die Vorderseite einer 3 mm dicken Kupferplatte schwarz lackiert und auf die Rückseite eine Heizmatte aufgeklebt (Abbildung 8, links). Ein etwa 10 mm dickes Passe-Partout wurde auf der Heizplatte platziert (Abbildung 8, Mitte), um das wertvolle Untersuchungsobjekt auf Distanz zur Heizplatte zu halten (Abbildung 8, rechts).

Mit einer Thermographiekamera, die vor dem Papier platziert wird, lassen sich qualitativ sehr hochwertige Aufnahmen von Wasserzeichen erzielen, wie in Abbildung 9 zu sehen ist. Hierbei handelt es sich um eine Handzeichnung in brauner Tinte, die vermutlich um 1665 von Jan Lievens¹⁰ hergestellt wurde (Döhring). Bevor das Wasserzeichen aufgezeichnet wurde, war nicht sicher, ob es sich bei der Zeichnung um ein Original aus der Schule Rembrandts handelt oder um eine Reproduktion aus dem 18. Jahrhundert (Laurentius). Das mit Hilfe der Infrarot-Technik gefundene Wasserzeichen bewies eindeutig, dass es sich bei der Handzeichnung tatsächlich um ein Original handelte.

⁹ X8011 Röntgen-Inspektionsgerät hergestellt durch die Firma Viscom.

¹⁰ Lievens, Jan, Bauerngehöft am Wasser, um 1665, Rohrfederzeichnung in brauner Tinte, $23,8 \times 36,4\ \text{cm}$; Herzog Anton Ulrich-Museum Braunschweig, Kupferstichkabinett, Inv. Nr. Z 103, (Copyright: Herzog Anton Ulrich-Museum, Braunschweig, Museumsfotographie von Claus Cordes).

Das Wasserzeichen in Abbildung 10 (links und Mitte) wurde mit dem Infrarot-System aus Abbildung 8 gewonnen. Die Handschrift, die schemenhaft im nahen Infrarot-Licht (Abbildung 3) noch sichtbar war, ist im MWIR-Spektrum von 3–5 μm völlig unsichtbar.

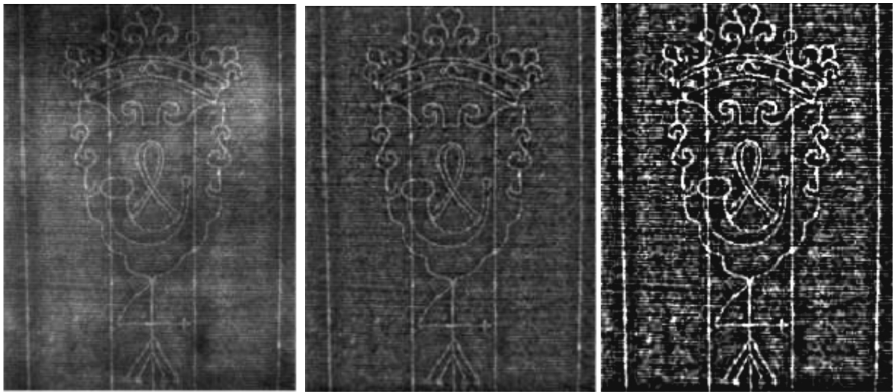
Falls die Sichtbarkeit eines Wasserzeichens sehr schlecht ist, lässt sich der Kontrast des vollständigen oder auch nur eines Teils des Wasserzeichens verbessern, so dass selbst kleinste Details herausgearbeitet werden können. Hierzu gehört auch ein Maßstab, der entweder nachträglich in das Bild z.B. mit dem Bildbearbeitungsprogramm Photoshop eingefügt wird (Abbildung 10, Mitte, linke Seite) oder auch ein Maßstab, der bei der Aufnahme gleich mit in dem Bild erscheint (Abbildung 10, Mitte, rechte Seite).

4. Auswerteverfahren

Verschiedenste Wasserzeichensammlungen von unterschiedlicher Herkunft, Detaillierung und unterschiedlichem Umfang sind heute in digitalisierter Form verfügbar. Dabei ist es eine besondere Leistung des »Bernstein-Projektes«, fünf verschiedene, online verfügbare Sammlungen von Wasserzeichen¹¹ unter einer einheitlichen Oberfläche den Nutzern verfügbar gemacht zu haben. Das »Bernstein-Portal« ermöglicht eine gleichzeitige Suche in den verschiedenen Datenbanken. Die großen Unterschiede in der Qualität der abgenommenen Wasserzeichenbelege (Zeichnung, Durchreibung, Durchlicht, Betaradiographie) und die in unterschiedlicher Detaillierung vorliegenden Informationen zu jedem Wasserzeichen erschweren allerdings eine Suche nach ähnlichen Wasserzeichen, da häufig zu wenig Details sichtbar sind. Ob also das richtige Wasserzeichen gefunden bzw. der richtige Schluss aus dem Vergleich des unbekannten Wasserzeichens mit den Wasserzeichen der Datenbanken gezogen werden kann, hängt ganz wesentlich von der Erfahrung der jeweiligen Wasserzeichenforscher und deren speziellen Kenntnissen über den Untersuchungsgegenstand ab.

Ein Ausweg aus dieser Unzulänglichkeit kann zum einen durch bessere Wiedergabe der Wasserzeichen mit dem oben vorgestellten neuartigen Verfahren zur Visualisierung mittels Thermographie gelingen, zum anderen bieten insbesondere bei einer verbesserten Bildqualität Verfahren der Mustererkennung Möglichkeiten zu einer leistungsfähigeren Unterstützung bei der Suche nach ähnlichen Wasserzeichen in Datenbanken. Im Folgenden werden systematisch verschiedene Ansätze zur Unterstützung der Suche vorgestellt. Diese Verfahren wurden bisher nur teilweise untersucht bzw. erprobt (Rauber u.a.; Wenger u.a.; Atanasiu; Hiary u.a.). Es ist die Aufgabe weiterer Forschung, geeignete Algorithmen anhand von Testdaten zu entwickeln und in Systeme zu integrieren, die ihre Eignung in Testszenarien beweisen müssen.

¹¹ Über das Bernstein-Portal sind die folgenden Sammlungen verfügbar: [NIKI](#) internationale Datenbank von Wasserzeichen und Papier genutzt für Drucke und Zeichnungen (1450–1800), [Piccard-Online](#) (überwiegend Archiv-Dokumente des 14.–16. Jh.), [WIES](#) – Wasserzeichen in Inkunabeln aus Spanien; [WILC](#) – Wasserzeichen in Inkunabeln aus den Benelux-Ländern, [WZMA](#) – Wasserzeichen des Mittelalters (Österreich, 14. –15. Jh.).



a) Original Thermographiebild

b) korrigiert

c) kontrastverstärkt

Abbildung 11. Darstellung des Wasserzeichens aus Abb. 2 mit einigen Bildbearbeitungsschritten.

Die Vorteile detailreicher, als Bild gespeicherter Wasserzeichen, wie sie z.B. durch thermographische Kameras erzeugt werden können, gegenüber Durchzeichnungen wurden bereits erörtert. Im Folgenden wird auf die Möglichkeit des Einsatzes von Mustererkennung bei den verschiedenen existierenden Repräsentationen von Wasserzeichen eingegangen.

In Abbildung 11 ist das Wasserzeichen des historischen Dokumentes aus Abbildung 2 dargestellt, wie es mit Hilfe der Thermographie völlig ohne den beidseitig aufgebrauchten Text deutlich sichtbar gemacht werden kann. Mit einigen wenigen Bildverarbeitungsschritten kann das Wasserzeichen deutlich hervorgehoben werden, so dass interaktiv durch einen Handschriftenbearbeiter oder -katalogisierer, später auch durch automatische Mustererkennung, wichtige Charakteristika erfasst und der Beschreibung des Wasserzeichens hinzugefügt werden können.

Abbildung 12 stellt einige Merkmale beispielhaft dar. Wichtig hierbei ist, dass zusätzlich zu den Motivinformationen, die der Benutzer aus Wasserzeichenrepertorien und -findbüchern erhält (z.B. Krone, Hifthorn) und dem Wasserzeichen zuordnet, auch automatisch weitere Merkmale und deren relative Positionen im Wasserzeichen, wie z.B. Kreise, Bögen, Ecken oder Geraden, und deren Eigenschaften (z.B. Größe und Ausrichtung) bestimmt und ebenfalls dem Wasserzeichen zugeordnet werden können. Anders als bisher können somit in einem ersten Schritt für ein unbekanntes Wasserzeichen Merkmale automatisch bestimmt werden, die dann in einem zweiten Schritt verwendet werden, um Wasserzeichen mit ähnlichen Merkmalen aus einer digitalen Datenbank herauszusuchen. In einem zukünftigen System könnte dieser

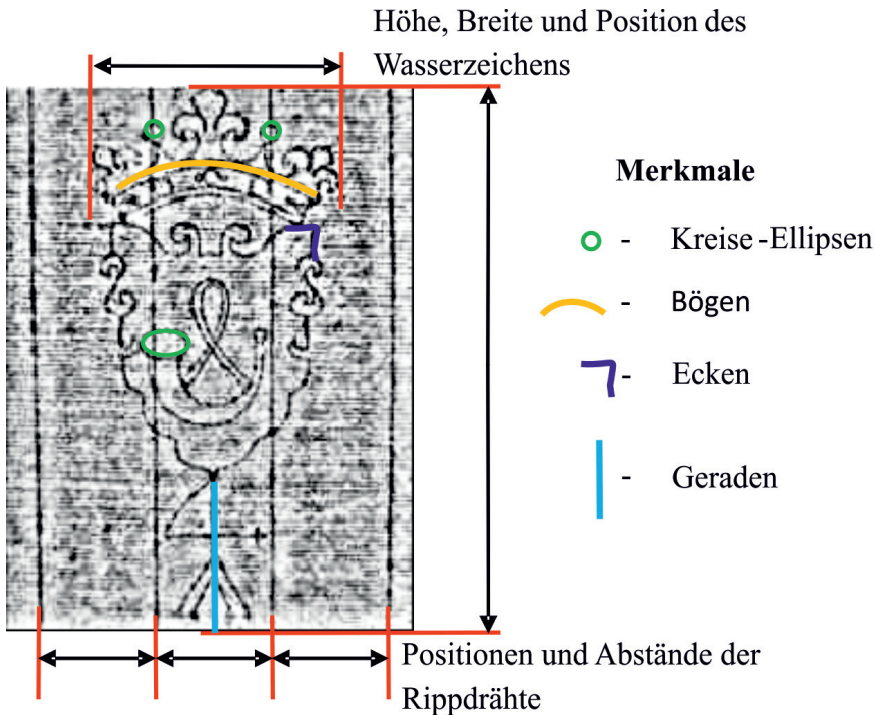


Abbildung 12. Verbessertes, invertiertes Wasserzeichenbild aus Abb. 11 mit beispielhafter Markierung einiger wichtiger Merkmale, die Wasserzeichen charakterisieren.

Suchvorgang (im Idealfall) völlig ohne Eingriff des Benutzers erfolgen. Er würde so ein erstes Suchergebnis liefern, welches unabhängig von externem Expertenwissen gewonnen wird. Erst im zweiten Schritt wird dann der Benutzer das gewonnene Ergebnis bewerten und die Suche weiter verfeinern. Die Erzeugung und die Bereitstellung von Wasserzeichensammlungen mit detaillierteren Merkmalen, die über die bisherigen verbalen und visuellen Beschreibungen hinausgehen, haben somit das Potenzial, die Verwendung von Wasserzeichen für die Erschließung historischer Schriften deutlich zu bereichern und die erzielbaren Ergebnisse zu verbessern.

Hier stellt sich die Frage, wie in der Übergangszeit, in der viele Wasserzeichensammlungen in unterschiedlicher Qualität, von Durchzeichnungen, Durchreibungen, Durchlichtvorlagen, Radiographien bis Thermographien existieren, eine bessere Unterstützung der Suche erfolgen kann. Neben der notwendigen Voraussetzung, die

Sammlungen gemeinsam recherchierbar zu machen, wie es mit dem »Bernstein-Projekt« begonnen wurde, muss für eine standardisierte verbale Beschreibung gesorgt werden, damit eine Suche effizienter werden kann. Auch im vorliegenden Fall heterogener Datensammlungen kann die Bildverarbeitung und Mustererkennung unterstützend wirken, da – insbesondere aus den Durchzeichnungen – ebenfalls Merkmale der Wasserzeichen automatisch extrahiert werden können. Tatsächlich ist die Merkmaldetektion in Zeichnungen zwar einfacher zu realisieren, dafür ist aber die gewonnene Aussagekraft weniger stark, weil die Zeichnung selbst ungenau sein kann und auf der subjektiven Wahrnehmung desjenigen beruht, der das Wasserzeichen auf diese Weise abgenommen hat. So können etwa Details in der Zeichnung nicht notwendig identisch mit Details im Wasserzeichen sein. Hier wird aufgrund der beschriebenen Belegsituation vermutlich nur eine grobe Einordnung möglich sein.

Die hier vorgestellte Möglichkeit, mit Hilfe einer Thermographiekamera schnell und effizient Wasserzeichen in hoher Qualität zu digitalisieren, bildet die Grundlage für die Entwicklung automatischer Mustererkennungsalgorithmen, die eine deutliche Verbesserung der Suchmöglichkeiten in Wasserzeichen-Datenbanken erlaubt. In Ergänzung zu den bereits abgeschlossenen Arbeiten im »Bernstein-Projekt« und zu dem zurzeit laufenden WZIS- Projekt sind hier die Grundlagen geschaffen, um künftige Wasserzeichensammlungen in hoher Qualität digital zu erstellen.

Bibliographie

- Atanasiu, Vlad. *Assessing paper original and quality through large-scale laid lines density measurements*. XXVI Congress of the Intl. Association of Paper Historians »Paper as medium of cultural heritage. The archaeology and conservation of paper«, 26 August – 6 September 2002, Rome/Verona. Vienna: Austrian Academy of Sciences, 2002.
 <<http://www.bernstein.oeaw.ac.at/ad751/atanasiu2002ad751en.pdf>>.
- Bernstein-Projekt. Vienna: Austrian Academy of Sciences, 2006–2009.
 <<http://www.bernstein.oeaw.ac.at/>>.
- Döhring, Thomas. *Aus Rembrandts Kreis. Die Zeichnungen des Braunschweiger Kupferstichkabinetts. Ausstellung im Herzog-Anton-Ulrich-Museum Braunschweig, 21. September – 17. Dezember 2006*. Petersberg: Imhof, 2006.
- Haidinger, Alois. »Datieren mittelalterlicher Handschriften mittels ihrer Wasserzeichen.« *Anzeiger der philosophisch-historischen Klasse der Österreichischen Akademie der Wissenschaften* 139 (2004): 5–30. doi:10.1553/anzeiger139s5.
- Hay, Louis. »Papiergeschichte, eine Hilfswissenschaft? Ein Ja und ein Nein.« *Papiergeschichte als Hilfswissenschaft*. 23. Kongreß der Internationalen Arbeitsgemeinschaft der Papierhistoriker, Leipzig 31. August 1996. Hg. René Teygeler. Leipzig: Deutsche Bücherei, 1996. 17–22.
- Hedges, Blair S. »A method for dating early books and prints using image analysis.« *Proceedings of the Royal Society A* 462 (2006): 3555–3573.

- Hiary, Hazem und Kia Ng. »Watermark: From Paper Texture to Digital Media.« *axmedis. Proceedings of the First International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'05)*. Los Alamitos: IEEE Computer Society, 2005. 261–264. doi:10.1109/AXMEDIS.2005.50.
- Ircam. 2009. <http://www.ircam.de/produkte/kamera_anzeigen_d.php?id=88>.
- Jaffé, Albert. »Zur Geschichte des Papieres und seiner Wasserzeichen. Eine kulturhistorische Skizze unter besonderer Berücksichtigung des Gebietes der Rheinpfalz.« *Pfälzische Heimatkunde* 3/4 (1930): 21–23.
- Kämmerer, Carmen. *Ars longa, vita brevis: Zeichenkunst im Alten Buch. Exlibris, Druckersignete und Wasserzeichen aus den Beständen der Stadtbibliothek Worms. Ausstellung der Stadtbibliothek Worms vom 23. Oktober 2008 – 22. November 2008*. Worms: Worms Verlag, 2008.
- Klinke, Thomas. »Die dritte Dimension. Methoden zur Feststellung technologischer Merkmale an historischen Künstlerpapieren und die Relevanz ihrer Erhebung.« *Journal of Paper Conservation* 4 (2009): 28–37.
- Kushel, Dan A. »Application of Transmitted Infrared Radiation to the Examination of Artifacts.« *Studies in Conservations* 30 (1985): 1–10.
- Kushel, Dan A. »Radiographic Methods Used in the Recording of Structure and Watermarks in Historic Papers.« *Fresh Woods and Pastures New, Seventeenth-Century Dutch Landscape Drawings from the Peck Collection*. Hrsg. Franklin W. Robinson und Sheldon Peck. Chapel Hill: Auckland Art Museum, 1999. 117–133.
- Lackner, Franz. »Barolus de Saxoferrato.« *Ochsenkopf und Meerjungfrau. Papiergeschichte und Wasserzeichen vom Mittelalter bis zur Neuzeit. Begleitbuch und Katalog zur Ausstellung des Landesarchivs Baden-Württemberg, Hauptstaatsarchiv Stuttgart und der Österreichischen Akademie der Wissenschaften, Kommission für Schrift- und Buchwesen des Mittelalters, Wien*. Stuttgart: Hauptstaatsarchiv und Wien: Österr. Akad. der Wiss., Komm. für Schrift- und Buchwesen des Mittelalters, 2009. 10–11.
- Laurentius, Theo. »Het onderzoek naar Rembrandt's papier. The investigation into Rembrandt's paper.« *Voelbaar papier: papierkunst in Nederland – Tactile paper*. Houten: Ekspress-Zo, 1996. 112–119.
- Maldague, Xavier P. V. *Theory and practice of infrared technology for non-destructive testing*. New York: Weinheim, 2001.
- Meinlschmidt, Peter. *Thermographie zur Detektion oberflächennaher Fehler – Grundlagen – Möglichkeiten – Grenzen – Anwendungen. SEF Abschlussbericht*. 2004.
- Meinlschmidt, Peter und Volker Märgner. *Erkennung von Dicken- und Dichteunterschieden in transparenten und halb-transparenten Materialien mittels Thermographie. Patentanmeldung DE 10 2008 016 195 B3*. 2008.
- Neuheuser, Hanns Peter, Volker Märgner und Peter Meinlschmidt. »Wasserzeichendarstellung mit Hilfe der Thermographie.« *ABI-Technik* 4 (2005): 266–278.
- Niki. Bad Neuenahr: Dutch University Institute for Art History / DiWe Media, 2010. <<http://www.wm-portal.net/niki/index.php>>.
- Piccard, Gerhard. »Die Wasserzeichenforschung als historische Hilfswissenschaft. Mit 25 Abbildungen.« *Archivalische Zeitschrift* 52 (1956): 62–115.

- Piccard, Gerhard. »Problematische Wasserzeichenforschung.« *Börsenblatt für den Deutschen Buchhandel* 61 (1965): 1546–1548.
- Piccard-Online. Stuttgart: Landesarchiv Baden-Württemberg.
<<http://www.piccard-online.de/start.php>>.
- Rauber, Christian et al. »Archival and retrieval for Large Image Databases: Application to an Historical Watermarks Archive.« *Proceedings of IEEE International Conference on Imaging Processing*, Lausanne (1996): 19–25.
- Thermosensorik. 2009. <http://www.thermosensorik.de/produkte_03.htm>.
- Weiß, Wiso. »Die Bedeutung der Wasserzeichenkunde für die Geschichtsforschung.« *Archivmitteilungen* 1 (1955): 18–25.
- Wolf, Christina. »Aufbau eines Informationssystems für Wasserzeichen in den DFG-Handschriftenzentren.« *Kodikologie und Paläographie im Digitalen Zeitalter*. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan. Norderstedt: BoD, 2009. 97–107.
<<http://kups.ub.uni-koeln.de/volltexte/2009/2963>>.
- Wenger, Emanuel et al. »A Digital Image Processing and database System for Watermarks in Medieval Manuscripts.« *Proceedings from the ichim01 meeting »Cultural Heritage and Technologies in the Third Millennium«*, September 3–7 2001, Milano, Italy. Band 2 (2001): 259–264. <http://www.archimuse.com/publishing/ichim01_vol2/wenger.pdf>.
- WIES: *Watermarks in Incunabula Printed in Spain*. Vienna: Austrian Academy of Sciences, 2007.
<<http://www.bernstein.oeaw.ac.at/databases/wies/index.html>>.
- WILC: *Watermarks in Incunabula printed in the Low Countries*. Koninklijke Bibliotheek - National library of the Netherlands, 2000–2007. <<http://www.kb.nl/bc/incun/watermerken-en.html>>.
- WZIS-Projekt: *Wasserzeichen-Informationssystem Deutschland*. Stuttgart: Landesarchiv Baden-Württemberg und Württembergische Landesbibliothek / München: Bayerische Staatsbibliothek / Leipzig: Universitätsbibliothek / Wien: Österreichische Akademie der Wissenschaften.
<<http://www.landesarchiv-bw.de/web/50960>>.
- WZMA: *Wasserzeichen des Mittelalters*. Wien: Österreichische Akademie der Wissenschaften, 2007. <<http://www.oeaw.ac.at/ksbm/wz/wzma2.htm>>.

Digitale Paläographie



Digital Palaeography

Teaching Manuscripts in the Digital Age

Peter A. Stokes

Abstract

This chapter reflects on the author's practical experience teaching palaeography in several different contexts at the start of the so-called "digital age". Material for manuscript-studies is becoming available at an enormous rate: perhaps most obvious are the results of the large-scale digitisation programmes which are making high-quality colour facsimiles of manuscripts available online to wide audiences. At the same time, Virtual Learning Environments provide new possibilities for teaching and learning, and many tools for research on manuscripts can also be used for teaching. Perhaps more fundamentally, however, it has often been noted that scholarship is changing as a result of digital tools, resources, and methods. What, then, of teaching? Should the teaching of manuscript studies also change along with the scholarly discipline, bringing the Digital Humanities into our classes on palaeography and codicology? To begin answering this question, and to suggest some pedagogical possibilities brought about by technology, the author's own experiences are discussed. Some limitations of technology for teaching are then considered, and some general remarks are then provided on the relationship between palaeography and Digital Humanities, two fields which are both fighting for recognition as full academic disciplines and not "mere" *Hilfswissenschaften*.

Zusammenfassung

Der Beitrag reflektiert die praktischen Erfahrungen des Autors, Paläographie in verschiedenen Kontexten zu Beginn des sogenannten "digitalen Zeitalters" zu lehren. Materialien zur Handschriftenkunde werden in großer Zahl verfügbar: Am offensichtlichsten sind die Ergebnisse groß angelegter Digitalisierungsprogramme, welche hochauflösende Faksimiledigitalisate von Handschriften einer weiteren Öffentlichkeit zugänglich machen. Gleichzeitig bieten virtuelle Lernumgebungen neue Möglichkeiten für Lehre und Lernen, und zahlreiche Werkzeuge der Handschriftenforschung können auch für die Lehre genutzt werden. Fundamental ist allerdings die oft zitierte Veränderung der Wissenschaft durch digitale Werkzeuge, Ressourcen und Methoden. Wie verhält es sich damit in der Lehre? Sollte sich auch das Lehren der Handschriftenkunde mit der wissenschaftlichen Disziplin ändern und die digitalen Geisteswissenschaften in Paläographie- und Kodikologiekurse einbringen? Eine Antwort auf diese Frage und

Anregungen für pädagogische Möglichkeiten durch Technologieeinsatz werden auf der Basis der Erfahrungen des Autors gegeben. Es werden die Grenzen des Einsatzes von Technik in der Lehre diskutiert und grundsätzliche Bemerkungen über das Verhältnis von Paläographie und digitalen Geisteswissenschaften gemacht, jenen zwei Feldern, die um die Anerkennung als volle akademische Fächer ringen und nicht "bloße" Hilfswissenschaften sein wollen.

1. Introduction

M.R. James once wrote "I cannot teach the art of assigning dates to manuscripts; I am even inclined to think that it cannot be taught" (Pfaff 103). Despite this claim, manuscript studies in general and palaeography in particular have often been taught in programmes of classics and medieval studies, at least at graduate level. However, these courses have tended to remain relatively constant in the way that they are taught and have not always taken advantage of new developments in the "digital age". This applies on several levels: on the one hand, there are many ways in which technology can improve the teaching of "traditional" palaeography, and this has certainly been done in some cases, although perhaps not as widely as one might like. However, there are other issues which are much less frequently discussed. It is often noted that the so-called "digital age" is transforming humanities scholarship, including traditional fields such as palaeography and manuscript studies (Vogeler; Stokes). If one accepts this, and indeed the present volume and its predecessor seem to demonstrate it, then it follows that teaching should change accordingly. Certainly students must learn the "traditional" skills which are central to manuscript studies and, many readers of this volume may argue, to much of the humanities in general. This author takes it as given that basic skills in handling original materials, in reading, transcribing, editing and understanding these objects is central to medieval and even much of modern studies. The question that remains is therefore twofold. First, how can digital tools be used to better teach traditional skills. Second, a question much less frequently raised, is how the teaching of traditional skills should or could itself change as a result: how and to what extent should digital content be explicitly introduced into the curriculum for the study of medieval manuscripts? It is this question that will be addressed here. The discussion will focus necessarily on the author's own experience and makes no claim to a comprehensive survey of all teaching on the subject; rather some practical experiences and broader theoretical considerations are offered in the hope that we can learn from our collective experiences and also reflect on how we teach a topic which has great potential for attracting students now more than ever, but the provision of which is undergoing significant and drastic change across Europe and beyond.

2. Medieval Manuscript Studies in the Digital Age

Medieval Manuscript Studies in the Digital Age (MMSDA) is a six-day training course for post-graduate (PhD) students in the United Kingdom. Its principle subject of study is “the analysis, description and editing of medieval manuscripts” for both print and digital output. It is funded by the UK Arts and Humanities Research Council (AHRC) under their Collaborative Research Training scheme. This scheme operates at three levels and was developed with an explicit purpose.

It enables institutions to offer such training to groups of students in several institutions where it is not possible or cost-effective to provide the training to students in just one department or institution. The expectation is that through this collaboration, an enhanced quality of training and student experience can be provided. (AHRC 2008 1)

MMSDA runs as a collaboration between four institutions. It is based at the Institute of English Studies, which is one of the Schools of Advanced Study in the University of London; the other collaborating institutions are the Department of Anglo-Saxon, Norse and Celtic in Cambridge, the Centre for Computing in Humanities at King’s College London, and the Warburg Institute which is another of the Schools of Advanced Study. Many further institutions are involved which are not formal collaborators, and a team of instructors and administrators makes the course run. In total, eighteen instructors and ten institutions have taken part in both of the first two years of the course.¹

As noted above, the stated subject of the course is “the analysis, description and editing of medieval manuscripts” for both print and digital output (MMSDA 2010). This reflects but perhaps does not make sufficiently explicit a central principle to the course which was incorporated into the planning from the very start: namely that both traditional and digital approaches should be taught together, with equal weight given to each, and with emphasis placed on how each one interacts with and enhances the other. It is this deep integration of digital and traditional, planned from the start as a fundamental principle rather than something which is added later or coincidental, which makes it different to any other which is known to this author. There are many intensive courses for teaching manuscript studies to postgraduate and other interested groups (although there are never enough).² Similarly, there are intensive courses for teaching digital methods, including XML markup and related technologies and approaches.³

¹ For a full list of current instructors and institutions see MMSDA 2010, “Schedule” and “Instructors”.

² A small sample of these includes the London Rare Book School and Palaeography Summer School at the Institute of English Studies in the University of London, and further examples in the UK are listed in Institute of English Studies and HOBOS.

³ For a list see EDU-SIG; further examples include the Digital Humanities Observatory Summer School in Dublin; Scholarly Codicological Research, Information & Palaeographical Tools (SCRIPTO), Friedrich

There are also courses in traditional palaeography which utilise digital methods, for which see further below, and courses which include some modules on manuscript studies and others on digital methods but without fully integrating the two.⁴ However, if the digital age truly has arrived, and if digital methods have become part of humanities scholarship; if digital approaches should be closely integrated with and informed by humanities scholarship; then surely it follows that teaching should reflect this and that both should be taught together as one integrated whole, rather than as two discrete parts as normally happens in practice. This is the next level in Digital Humanities and Humanities in the “digital age”.

How, then, did this course work in practice? The first three days are spent on “traditional” manuscript studies, with lectures on palaeography, codicology, art history, principles of cataloguing, provenance, and principles of editing. These classes include visits to libraries with significant collections of medieval manuscripts in Cambridge and London, during which the principles discussed in the morning classes are then worked through with real examples. In this way, a firm theoretical basis was established, along with some experience in practical applications (and, we hope, some exposure to the difficulties which theories inevitably encounter when put into practice). The next three days then focus on these same principles but applied to the digital realm. How does one catalogue in a digital format? How does one present a catalogue online? What of an edition? How do the principles and practices change when applied to a digital format? These questions are all addressed, both through discussion and implicitly through practice as the students produce their own sample catalogue-entries, transcriptions, and edited passages. In order to aid these questions, and to provide links across the different components of the course, we have made extensive use of Parker on the Web and the collection of Corpus Christi College in Cambridge, as well as the generosity of the library’s staff. We are fortunate enough to have access to a wide range of material which we could draw on for teaching: the original manuscript, M.R. James’ printed catalogue of 1909–12 (a scan of which is now freely available in PDF format from the library website), free access to the new digital catalogue provided by Parker on the Web, and high-quality images of the manuscript provided by the college (cf Gillespie 2010).⁵ After lectures on cataloguing, palaeography and art history, we take the students to the Parker Library at Corpus Christi College, and among many manuscripts we show them manuscript 422, the “Red Book of Darley”. The students then produce brief electronic

Alexander University Erlangen-Nuremberg; TEI @ Oxford Summer School at the University of Oxford; and this author’s own course on Digital Publishing for the London Rare Book School in the Institute of English Studies, University of London.

⁴ Examples include SCRIPTO and (from July 2010) the London Rare Book School, as well as numerous postgraduate Masters courses or equivalent on material culture.

⁵ The author, on behalf of the MMSDA team, wishes to thank Stanford University Libraries and Harrassowitz for granting access to Parker on the Web during the course.

catalogue descriptions of this manuscript in TEI-compliant XML, based on what they have seen in the library but also supplemented by the digital facsimile at Parker on the Web. We then take M.R. James's description and compare that with the new Parker on the Web to see the advantages and disadvantages of online *versus* print cataloguing. The students also electronically annotate the digital image of one of the more complex pages from the manuscript and use this with the Image Markup Tool (IMT; see Holmes) to create their own web page of a small digital edition which integrates image and text. This is quite a different use of Parker on the Web from that which the creators presumably intended: on the one hand, the digitised manuscript functions as material for the students, but the catalogue description also serves as a model for them to imitate in the first instance and as the object of their analysis in the second. Having both the print and online catalogues provides further material for discussion, allowing the students to compare the strengths and weaknesses of each. Without both the images and the two detailed catalogues online, the whole exercise would not have been possible.

The success of this approach has been significant. Enrolments and feedback provide the first evidence of this: with more than three applicants for every position in the first year, and about thirty applicants for twenty places in the second and third years, even though places were limited to PhD students registered at UK institutions. Feedback from students, their supervisors, and instructors on the course has been unanimously very positive indeed. All but two of the instructors and institutions from the first year participated again in 2010, and most also in 2011; of the two who did not in 2010, one is an instructor who has since changed field and moved country, and the other is an institution which is already heavily burdened by prior commitments but which is participating again in 2011. In addition to these quantitative measures of success are some others which are more subjective but which are perhaps of greater significance. When the students apply to the course, they are asked to write a brief statement outlining what they expect to gain from it: almost without exception, they focus on the manuscripts, the need for experience in handling original artefacts, and their lack of training in skills such as palaeography, editing, cataloguing and (much less often) art-history. Very few mention the "digital" element, and those who do usually show little awareness of its significance or the issues involved. In contrast, feedback after the course generally reveals a significant increase in understanding issues of digital production and consumption, including the value of standards such as TEI XML, greater critical awareness in the use of online resources, and awareness of the interrelation between digital and non-digital practices. This increased awareness is difficult to quantify, not least because we have no formal test either before or after the course, so one anecdote must serve to illustrate this point. When the course first ran in 2009, the visits to the libraries coincided with the traditional lectures and such in the first three days: the course was therefore itself divided into two clear parts, the "non-digital" and the "digital". In this respect we ourselves were guilty of the false

division just criticised in the first section of this chapter. It was the students who pointed this out, as one noted in feedback at the end of the course that it would have been useful to schedule a final library visit after they had gained some exposure to digital methods so that they could bring this experience back to the manuscripts themselves, and this suggestion was incorporated into the course for 2010.

3. Teaching Undergraduates using Digital Resources

In addition to the MMSDA course, this author's other experience includes teaching undergraduate students in a Palaeography and Codicology course in the Department of Anglo-Saxon Norse and Celtic in the University of Cambridge, and one on Book History for the School of Historical Studies in the University of Leicester. The former is aimed primarily at first and second-year undergraduates but usually includes one or two MPhil students and is occasionally audited by PhD students. The latter is for second-year undergraduates. Both courses have proven very popular: enrolments at the one in Cambridge peaked at about thirty students, and the one in Leicester has only been offered once but had a similar number of students. The course in Cambridge initially involved four or five contact-hours a week, but these hours were reduced because of the costs involved and so online teaching has become more important as a result. Learning-outcomes for both courses included a sense of the material culture of the early Middle Ages, some awareness of the survival (or lack thereof) for medieval documents and manuscripts, and other similar aspects that are relatively unsuited to current online teaching for palaeography and manuscript studies for reasons that have already been discussed elsewhere and will be again shortly here. Further outcomes of the Cambridge course were the ability to describe, date, localise, and transcribe scribal hands. These skills involve training rather than teaching: they require students to invest a significant amount of time practicing, preferably assisted by feedback from an instructor of some sort, and for this reason the course is in many ways closer pedagogically to language teaching than to history. This meant on the one hand that the reduced contact hours required finding new ways of providing the extended training and feedback, but also that online methods could be applied more readily than to the material aspects of book history, although this holds only with some important caveats that will be discussed below.

Both Cambridge and Leicester now use Virtual Learning Environments (or VLEs): the former Sakai and the latter Blackboard. Both are very similar in functionality, perhaps the biggest difference being that Blackboard is a proprietary system developed by a commercial firm, whereas Sakai is free and open-source, developed by an international community (Sakai Project; Earhart). The principle advantages of VLEs are obvious, insofar as both systems provide an easy way for instructors to add images or links

to images for students. Instructors can create online assignments, where students are told to go to a particular image of a particular manuscript on-line, to transcribe a certain number of lines, and to discuss features of the manuscript page (or entire manuscript); the transcription and discussion can be submitted through the website, and the instructor can then correct it and give feedback. One can also illustrate a single page of a manuscript in class, giving the students a black-and-white printout for them to annotate during the class but providing them with a link to a full-colour, high-quality image to view in their own time.⁶

This use of digital images was itself an obvious but very significant improvement to the course which initially involved plates in books being photocopied in black and white onto overhead transparencies: many students expressed their appreciation of the greatly improved quality of the presentations in class (cf. Duggan; Twycross; Kamp; Gillespie). It also allowed other possibilities, however. Rather than using simple static images, as in an overhead transparency, slide, or even PowerPoint presentation, high-quality digital images can be manipulated “live” during the class. One can therefore project an image of an entire manuscript page to discuss topics such as *mise en page* but can then zoom in to different regions to illustrate details of script, decoration, glosses and the like, something which this author and others have found invaluable in the classroom (cf. Duggan 156–7) or indeed in conference presentations. As well as providing a more arresting class, this also conveys better the relationship between the details and the whole.

All of this is made easier by digital resources, but none of it is particularly exceptional or even very different from what was done previously. More interesting teaching becomes possible when complete manuscripts are available on-line: in these cases the instructor might discuss one page during a lecture, but the students can then go to the complete facsimile and see that page in its larger context: they therefore gain access not only to a very good image but also to the rest of the book, which allows them to supplement the lecture material with their own investigation. Exercises can work the other way, too. For example, a class on liturgical manuscripts can involve searching catalogues for litanies and then comparing the facsimiles of litanies from different manuscripts. Patterns of survival can be investigated by searching catalogues for different types of book from different dates and different locations and seeing how these change. The resources for these sorts of classes are available now, and with a little creativity students can be given all manner of questions which they can then investigate using digital resources in a form of active learning that has been used very effectively with the *Oxford English Dictionary*, to name just one example (Simpson; Bunting and Stevens). With the advent of so-called “Web 2.0” this could easily be taken

⁶ In Cambridge, the department has a site licence for Bernard Muir’s *Ductus* programme which has very high-quality images and transcriptions, although these are just single pages of manuscripts (Muir and Kennedy 2007; Muir 2009 139–40).

even further. One important aspect of “Web 2.0” is user-created content: that is, people can now create their own web pages by aggregating content from different sources, as well as blogging, commenting on each other’s pages, and so on. The significance of this has not been lost in educational circles, and resources such as Google Groups are already being used in teaching: indeed, this is one of the principle purposes of resources such as Sakai and Blackboard (Mahony; cf Kamp). Content can easily be presented in new ways, too, such as Simile Timemaps which allow one to rapidly create web pages that integrate data, maps and timelines that can be linked back to online facsimiles and catalogues (*Timemap*). For example, in this author’s experience at Cambridge, students tended to become too tied down in the *minutiae* of letter-forms and quickly lost the larger-scale overview of chronological and geographical developments. This could then be ameliorated by providing a map on which the manuscripts are plotted by their (presumed) place of origin, and which are also marked on the accompanying timeline: they can therefore manipulate both the timeline and the map to gain a sense of the “bigger picture”. The map and timeline both include links back to the online facsimiles of the various manuscripts, thereby allowing students to go back to the details. These timelines and maps can read directly from XML files, so it would be very easy to create pages that read their data directly from the online catalogues if this material were made freely available, although in practice this is rarely the case. In his own teaching this author has only produced a few very crude examples, and has come nowhere near using their full potential, but this potential seems clear and should be exercised more (Mahony). Indeed, many of the uses of digital resources for research in manuscript studies that this author has discussed elsewhere (Stokes) can be applied equally usefully to teaching; virtual light-boxes are one obvious example (*Online Gallery*), as are annotating images and sharing annotations (OCVE), tools to aid transcription such as one developed by Jim Ginther which has already been used to teach palaeography (Ginther; Gillespie), a Virtual Research Environment for the study of documents and manuscripts (Bowman et al.), and many others.⁷

4. Some Limits of Digital Teaching

The discussion so far has outlined some possibilities for teaching which have already been put into practice. Many other possibilities exist, and this chapter makes no attempt to be complete in any way, but it is hoped that some of these may help to show what can be done. However, there are also many aspects of manuscript studies that cannot be taught easily or at all with existing online resources. Once again a full discussion of

⁷ Further examples can be found in the first volume of *Codicology and Palaeography in the Digital Age* (KPDZ 1), particularly the contribution by Kamp and by Cartelli and Palma.

these is beyond the scope of this chapter, and so two of them will now be discussed in some depth.

4.1. Codicology and the Materiality of the Book

One relatively obvious limit to computer-aided teaching is codicology and what has been termed “phenomenology” or the “materiality of the book”. Focussing on images reduces or removes other aspects of manuscript studies, including not only practical concerns such as how to prepare quire-diagrams and how to handle original artefacts safely, but also phenomenological issues such as their physical size and weight, or indeed how they feel, sound and smell. These last aspects are often ignored by scholars as well as students, but their importance has been stressed by some, and it has also been argued that, in the Middle Ages, the physical feel of a manuscript is almost as important as its appearance – that touch was close in significance to sight – and that digitisation increases further the modern privileging of sight over all other forms of acquiring knowledge (Treharne). Sight as the highest sense is clearly not a modern development, as it dates to Aristotle if not earlier, and we will never be able to reproduce the complete experience of a medieval reader, but it does remain that digital images present only the visual aspect of a manuscript, without giving much or any sense of its size and weight and often minimal sense of its format. However, size, weight and format are important parts of a manuscript, and even more so of a roll. The size of a manuscript tells us much about its function and status: a pocket-gospel was probably made for personal use, whereas a large-format bible may be an assertion of wealth and power. The enormous size and weight of a pipe roll is almost impossible to convey in digital format, but its size is also perhaps an assertion of authority and certainly tells us something about how the roll could and could not have been used. This author is probably not alone in badly underestimating the difficulty in handling a roll of this size, and therefore the significant amount of time required to check even one small detail in it.

These issues are well known, and are also becoming increasingly difficult to teach as libraries restrict access to their materials and are being squeezed more and more by funding cuts.⁸ A certain amount of this can be overcome by careful use of images and video, as demonstrated by Bernard Muir’s *Making of a Medieval Manuscript* (2007; Muir 2009 142): pedagogical aids such as this convey the structure and process of book-production perhaps even more effectively than a complete manuscript can, although this is only one part of codicological training. A good deal can be achieved with mockups, too, and this has been used in both the Cambridge undergraduate and MMSDA courses

⁸ As just one example of this, until recently conservationists at the British Library would train groups of students in the proper handling of books and rolls, but it has now become extremely difficult to arrange such sessions due to organisational and funding changes within the library.

with model books that have been produced by the Cambridge Conservation Consortium. Nevertheless, in this author's experience, libraries are still generally open to hosting groups of students if given appropriate warning and compliance with library procedures, and many universities also hold collections of fragments and manuscripts which can be used for teaching. Both the post-graduate MMSDA and the undergraduate course in Cambridge involve library visits, and almost all libraries have agreed to these without hesitation.⁹ Even if manuscript collections of the size and quality of Cambridge and London are not available, much can still be done with a little thought. Many universities have teaching collections of cheap manuscripts or manuscript fragments, and such a collection can be compiled very easily and with minimal budget even just by using eBay. Students can be asked to create mockups of manuscripts based on quire diagrams, or asked to think about the phenomenology of their own books: how is a paperback novel different from a glossy hard-cover coffee-table book, for example. Images projected during lectures can (and should) be accompanied by an indication from the lecturer of the approximate size and weight of the manuscript, and students should always be reminded to ask themselves if they know these details when they look at photographs or digital images. Such reminders and other indicators could conceivably be put into online teaching, but it remains that existing digital approaches are unable to deal with this satisfactorily.

4.2. The Problem of Transcription

One of the basic teaching outcomes in palaeography is transcription: that is, students should acquire the skill of reading and accurately transcribing original manuscripts, usually in a range of different scripts. This, like all other skills, is something that requires practice: I would argue that it can be taught, certainly, and that it does require guidance, but also that it requires each student to devote a relatively large amount of time in front of manuscripts or facsimiles gaining "hands-on" experience. As pressure on teaching increases, with larger classes and reduced time for teaching, it has become increasingly difficult to provide students with the time and attention that is necessary.¹⁰ The only way of maintaining the necessary skills is to require students to work more on their own, completing exercises in their own time which the instructor must then try to correct and give feedback on as best he or she can. The natural question thus arises

⁹ The MMSDA course involves visits to Corpus Christi College, Trinity College, and St John's College libraries in Cambridge; and Lambeth Palace, the Wellcome Institute, and the University of London Senate House libraries in London. The Cambridge undergraduate course has involved visits to the Cambridge University Library, Corpus Christi College, Cambridge, and the conservation rooms at the British Library.

¹⁰ To cite this author's experience, when he began teaching undergraduate students he was allocated five hours a week for sixteen weeks a year for discussion classes, one of which was dedicated to transcription. This has now been reduced to one hour a week for four weeks a year, with no time dedicated to transcription per se. Compare also the same trend observed by Ganz (1997).

whether digital methods can be used to supplement traditional classroom teaching and to make this process easier and less time-consuming.

To this end, and also for the benefit of those who do not have access to a formal course and wish to teach themselves, a relatively large number of online exercises has emerged (Vorholt *et al.*). In general these tend not to consider the dating or localisation of manuscripts much, if at all, although there are some exceptions: perhaps this is because people tend to accept James' principles that these skills cannot be taught, or perhaps the feeling is rather that they require the direct assistance of an instructor. What online courses very often do offer is exercises in transcription. In principle this is a valuable and logical approach. Facsimiles of manuscripts can be presented on the screen, students can enter their transcriptions, receiving a range of different types of help if required, and they can be given immediate feedback. Errors in transcription can be highlighted immediately, students can learn from their mistakes, and so they can build up the necessary skills in their own time, thereby supplementing classroom instruction and allowing the instructor to focus on particular areas where the students are having the most trouble.

This, at any rate, is the ideal, but it has a flaw in the very principle of its construction. The model usually assumes that every facsimile has one and only one correct transcription: the instructor, or the person who develops the teaching module, enters this one transcription, and all students' responses are checked against that one; any deviation from this is automatically marked incorrect. This model, then, assumes that transcription is entirely objective and therefore invariant for different transcribers. In practice, however, as has been repeatedly shown, transcription is not objective but requires interpretation, and there is often more than one possible reading for any given passage (Parkes xxix–xxx; *Computers and Old English Concordances* 88; Page 79; Robinson and Solopova 19; Robinson 43–44; Walsh; Pierazzo). Granted the extent of this interpretation varies: one can easily find many set scripts for which complete agreement on the readings can easily be achieved. However, as the level of cursiveness increases, or as the minims become more indistinguishable; as the level of abbreviation increases, and as scribes use more and more individual features, so the degree of ambiguity increases, and so the number of possible readings increases with it. Some of these possibilities will be eliminated by the principles of transcription which the students should have been given in advance—for example, whether or not to expand abbreviations, how to treat *u/v* and *i/j*, and so on—but even the most detailed guidelines cannot encompass all possibilities. These ambiguities in manuscript readings may also be irrelevant if one is editing a work, since the grammar may demand only one of the possible forms, and part of the editor's responsibility is to resolve these ambiguities according to the editorial principles and requirements of the edition. However, it may not be reasonable or even possible for a student to determine which of the various transcriptions is required by the sense. Besides, students are repeatedly told that the goal of transcription is to

reproduce what the scribe wrote, not what the sense requires: by definition, a diplomatic transcription must reproduce all scribal errors. When Malcolm Parkes (xxix–xxx) has noted that he cannot resolve the ambiguities between otiose strokes and abbreviation strokes in fifteenth century cursive, how can we expect the student to guess which reading the computer demands? In some online transcription exercises, the students are not even given any context, but are simply presented with the image of a single word, sometimes in a very cursive script with many possible readings. In a classroom, this is not a significant difficulty, since the instructor can adapt accordingly, perhaps explaining that the student's response is perfectly reasonable, and perhaps correct, but that the context makes an alternative answer more likely. However, with the online teaching that this author has seen, students are simply told that they are wrong.¹¹ In some cases the students are given no constructive feedback at all, but simply left to guess why they were wrong: this could be due to genuinely misreading a word, but it may also be rather because of a different interpretation of an ambiguous reading. Telling students that they are simply "wrong" is often counterproductive at best, and doing so without any explanation of why they were wrong or how to improve can be outright destructive. In the case of online teaching, the most likely result is for students to become discouraged and give up.

How are we to resolve this difficulty? One possibility is to present students with model transcriptions and ask them to correct their own work; this is already followed in most online courses,¹² but there are disadvantages with such an approach. First, it depends on the students taking the time and care to correct their work accurately. It also limits the amount they can learn from the exercise, since they lack the feedback of an instructor who may be able to recognise patterns in the students' errors and suggest ways of improving. It also fails to address the issue of variant answers, since the students will again only have access to one possible "correct" answer and may not be able to recognise that some errors are more venial than others, so to speak. A third possibility is for the instructors to correct the transcriptions by hand: this again has been followed in practice (Twycross 279–80; Muir 142; Gillespie; see above) and is probably the best in pedagogical terms, but it requires a significant investment from the instructors which they may not be in a position to make. A fourth possibility is to think more widely in the way we design our online transcription exercises, and in particular to look at teaching in other, related subject-areas. In general, the model for palaeography, implied but rarely made explicit, is that of foreign languages: just as languages require regular practice with a trained instructor, so do palaeographical skills, and just as machines have long been used to assist in the teaching of languages, so technology has been used to supplement palaeographical teaching, first with photography and now with

¹¹ Examples include Burghart; TNA "The Ducking Stool Game"; *Medieval Palaeography*; and Tillotson.

¹² Examples are Paläographie Online; Toureille; De Brún; EHOC; Scriptorium; TNA (except for the "Ducking Stool Game"; Muir 2009 141–2; Kamp 115–6.

interactive learning environments. But one would not normally type a translation of a text into a field and expect the computer to correct it. Instead, language software provides a wide range of alternative methods, most of which are more or less variations on multiple choice. So why should we expect transcription to be any different?

5. Final Remarks

This discussion has given some suggestions as to how palaeography and manuscript studies can be taught in the so-called “digital age”, as well as some limitations in the same. These limitations are significant and are not to be dismissed lightly; it is also not obvious how they can be overcome at least in the short term, and this is one of several reasons why digital technology should only ever supplement rather than replace teaching with a live human instructor. Nevertheless, the advantages which arise from supplementing teaching in this way are significant and should not be passed over, either through fear of the limitations or perceived lack of technical ability. Some of the suggestions given here do require some skill in computing but this is reducing rapidly as more and more tools and resources become available. The teaching of palaeography has long been a somewhat marginal and threatened activity (Lowe 580; Brown 378; Ganz 1990 and 1997), and this is particularly evident at the time of writing when the Chair of Palaeography at King’s College London has just been closed down, and the prospect looms of *all* government funding being cut for the teaching of Humanities subjects in UK universities.¹³ Nevertheless, student interest in the field remains high, as demonstrated by enrolment levels in the courses described here, and it is perhaps easier now than it has ever been before to attract students, since even universities without significant manuscript holdings can use the wealth of online resources: students can be tempted by the prospect of leafing through online facsimiles of the Lindisfarne Gospels, the Sherborne Missal, the Sforza Hours, or the Baybars’ Qur’an, to list just some offerings from just one freely accessible resource (*Turning the Pages*). In addition, however, it must be remembered that the digital is now part of humanities scholarship: it is no longer an adjunct field but is an integral part of it. Both palaeography and digital humanities have fought for recognition as valid fields of research rather than “mere” *Hilfswissenschaften* (Brown 361; Ganz 1990 17 and 1997 4; IATH; Terras 2010a), and neither should be taught without due recognition of the other. Students of medieval studies, particularly post-graduate students looking towards post-doctoral and academic careers, need a thorough grounding in the theory and practice both of digital humanities and of manuscript studies. Indeed very many post-doctoral positions being advertised

¹³ Discussion of this is far too voluminous to cite in full. Perhaps the highest-profile responses to the closure of the chair are Beard 2010, CIPL 2010, and Morgan 2010; see also Palaeography Working Group 2010. For the cuts to education in the Humanities, and the implications for Digital Humanities in particular, see especially Prescott 2010 and Terras 2010b.

at the time of writing require or prefer skills and experience in both areas, skills and experience which most post-graduate training does not provide. This is the rationale for MMSDA, both theoretical and practical, and the response seems to demonstrate that it is well founded.¹⁴

Bibliography

- AHRC: *Collaborative Research Training Scheme: Guidance Notes*. Arts and Humanities Research Council, 2008.
- Beard, Mary. "A Don's Life: University Cuts, Redundancies - and Bye-Bye Palaeography at King's College London." *Times Literary Supplement*, January 28 (2010). <<http://www.guardian.co.uk/education/2010/feb/09/writing-off-last-palaeographer-university>>.
- Blackboard. 1997–2010. <<http://www.blackboard.com/>>.
- Bowman, Alan K., et al. "A Virtual Research Environment for the Study of Documents and Manuscripts." *Digital Research in the Study of Classical Antiquity*. Eds. Gabriel Bodard and Simon Mahony. Farnham: Ashgate, 2010. 87–103.
- Brown, T. Julian. "Latin Palaeography Since Traube." *Transactions of the Cambridge Bibliographical Society*, 3 (1959–63): 361–81. Reprinted in *A Palaeographer's View: The Selected Writings of Julian Brown*. Eds. Janet Bately, Michelle P. Brown and Jane Roberts. London: Harvey-Miller, 1993. 17–37.
- De Brún, Pádraig. *Leathanaigh shamplacha ó lámhscríbhinní Gaeilge i Leabharlann na hOllscoile, Corcaigh*. Cork: University of Cork, 2002. <<http://www.ucc.ie/faculties/celtic/lss/>>.
- Bunting, David and Jessica Stevens. "Learning Resources." *Oxford English Dictionary*. Oxford: Oxford University Press, 2009. <<http://www.oed.com/learning/>>.
- [Burghart, Marjorie.] *Comptes des Chatellenies Savoyardes: Exercices de paléographie*. [Région Rhône-Alpes, 2010.] <<http://paleographie.castellanie.net/>>.
- CIPL: Comité international paléographie latine. "Abolition de la chaire de paléographie à King's College London." [Paris: Institut de recherche et d'histoire des textes, 2010.] <<http://www.palaeographia.org/cipl/actu/paleoatkins.htm>>.
- Computers and Old English Concordances*. Eds. Angus Cameron, Roberta Frank and John Leylerle. Toronto: University of Toronto Press, 1970.
- Duggan, Mary. "Teaching Manuscripts from the Web." *Literary and Linguistic Computing* 14:2 (1999): 151–60.
- Earhart, Amy. "Delivering Course Management Technology: An English Department Evaluates Open Source and For-Profit Course Management Systems." *Digital Humanities 2006*:

¹⁴ A course as large and complex as MMSDA requires the help and generosity of very many people, most of whom cannot be named here. The author wishes to thank everyone involved in the planning and running, particularly the instructors (for which see MMSDA: "Instructors"); staff and faculty at the Institute of English Studies, particularly Michelle Brown, Warwick Gould and Jon Millington; and most of all the core team of Hanna Vorholt, Elena Pierazzo and Arianna Ciula.

- Conference Abstracts*. Eds. Chengan Sun, Sabrina Menasri and Jérémy Ventura. Paris: Centre Cultures Anglophones et Technologies de l'Information, 2006. 302–3.
- EDU-SIG. TEI Education SIG. *Short TEI Training Courses*. 2004.
 <http://www.tei-c.org/Activities/SIG/Education/short_courses.xml>.
- EHOC. *English Handwriting, 1500–1700: An Online Course*. Cambridge: University of Cambridge, 2008. <<http://www.english.cam.ac.uk/ceres/ehoc/>>.
- Ganz, David. “Editorial Palaeography”: One Teacher’s Suggestions.” *Gazette du Livre Médiéval*, 16 (1990): 17–20.
- Ganz, David. *Latin Palaeography since Bischoff*. Transcript of Inaugural Lecture. London: King’s College London, 1997.
 <<http://www.kcl.ac.uk/content/1/c6/04/42/91/inaugural-lecture-1997.pdf>>.
- Gillespie, Alexandra. “Proposal – Use Case for Parker on the Web: Palaeography and Codicology Course.” Presentation to the Digital Manuscript Uses and Interoperation Workshop, Paris, 14–15 January 2010. [Abstract and slides:] <<http://lib.stanford.edu/digital-manuscript-uses-and-interoperation-public-site/abstracts-and-presentations>>.
- Ginther, James. “The *Electronic Norman Anonymous Project*: A Case for the Possibility and Limits of Interoperability.” Presentation to the Digital Manuscript Uses and Interoperation Workshop, Paris, 14–15 January 2010. [Abstract and slides:] <<http://lib.stanford.edu/digital-manuscript-uses-and-interoperation-public-site/abstracts-and-presentations>>.
- Holmes, Martin. *The Image Markup Tool*. Victoria: University of Victoria, 2010.
 <http://tapor.uvic.ca/~mholmes/image_markup/>.
- HOBO: *History of the Book Online*. Oxford: Faculty of English.
 <<http://www.english.ox.ac.uk/hobo/>>.
- IATH. “Is Humanities Computing an Academic Discipline?” An Interdisciplinary Seminar.” Institute for Advanced Technology in the Humanities. Charlottesville (VA): University of Virginia 1999. <<http://www.iath.virginia.edu/hcs/>>.
- Institute of English Studies. *National Research Training Portal*. London: Institute of English Studies, University of London, 2010. <http://ies.sas.ac.uk/nrts/events_search.php>.
- James, Montague Rhodes. *A Descriptive Catalogue of the Manuscripts in the Library of Corpus Christi College, Cambridge*, 2 vols. Cambridge: Cambridge University Press, 1909–1912. Online:
 <<http://sul-derivatives.stanford.edu/derivative?CSNID=88881001&mediaType=application/pdf>>, (vol. 1),
 <<http://sul-derivatives.stanford.edu/derivative?CSNID=88881002&mediaType=application/pdf>>, (vol. 2).
- Kamp, Silke. “Handschriften lesen lernen im digitalen Zeitalter.” *KPDZ 1*. 111–122.
- KPDZ 1: *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Eds. Malte Rehbein, Patrick Sahle and Torsten Schaßan. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. Online: <urn:nbn:de:hbz:38-29393>, <<http://kups.ub.uni-koeln.de/volltexte/2009/2939/>>.
- Lowe, Elias Avery. “The Ambrosiana of Milan and the Experiences of a Palaeographer.” *Palaeographical papers, 1907–1965*. Ed. Ludwig Bieler. Oxford: Clarendon Press, 1972. Vol. II: 575–90.

- Mahony, Simon. "An Interdisciplinary Perspective on Building Learning Communities Within the Digital Humanities." *Digital Humanities 2008: Book of Abstracts*. Eds. Lisa Lena Opas-Hänninen et al. Oulu: University of Oulu, 2008. 149–52.
- Medieval Palaeography*. Medieval Research Centre, University of Leicester. <<http://paleo.anglo-norman.org/medfram.html>>.
- MMSDA: *Medieval Manuscript Studies in the Digital Age*. London: Institute of English Studies, University of London, 2010. <<http://ies.sas.ac.uk/study/mmsda>>.
- Morgan, John. "Writing was on the Wall for Palaeography Chair." *Times Higher Education*, 21 October 2010. <<http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=413940&c=1>>.
- Muir, Bernard James and Nick Kennedy. *Ductus: A History of Handwriting*. CD-ROM, Version 2.0. Melbourne: E-Vellum, 2007.
- Muir, Bernard James. *The Making of a Medieval Manuscript*. Melbourne: E-Vellum, 2007.
- Muir, Bernard James. "Innovations in Analyzing Manuscript Images and Using them in Digital Scholarly Publications." *KPDZ* 1. 135–44.
- OCVE: *Online Chopin Variorum Edition*. London: King's College London, 2010. <<http://www.ocve.org.uk/>>.
- Online Gallery*. London: British Library, 2010. <<http://www.bl.uk/onlinegallery>>.
- Page, Raymond I. "On the Feasibility of a Corpus of Anglo-Saxon Glosses: The View from the Library." In *Anglo-Saxon Glossography: Papers read at the International Conference held in the Koninklijke Academie voor Wetenschappen Letteren en Schone Kunsten van België, Brussels, 8 and 9 September 1986*. Ed. Rene Derolez. Brussels: Paleis der Academiën, 1992. 77–96.
- Palaeography Working Group. *Final Report made to the Chair of the Executive Board of the School of Arts and Humanities*. King's College London, 30 June 2010. <<http://www.kcl.ac.uk/content/1/c6/07/64/51/ThePalaeographyWorkingGrouppaper.pdf>>.
- Paläographie Online. Von der römischen Antike bis zum Ende des Handschriftenzeitalters (1.–16. Jahrhundert)*. Eds. Peter Orth and Georg Vogeler. Erlangen, München: LMU, FAU, VHB 2005–2010. <<http://www.palaeographie-online.de/>>.
- Parkes, Malcolm Beckwith. *English Cursive Bookhands 1250–1500*. Oxford: Clarendon Press, 1969.
- Pfaff, Richard W. "M. R. James on the Cataloguing of Manuscripts: A Draft Essay of 1906." *Scriptorium*, 31 (1977): 103–118.
- Pierazzo, Elena. "The Limits of Representation: A Rationale of Digital Documentary Editions." Forthcoming.
- Prescott, Andrew et al. "Digital Humanities and the Cuts." *Humanist* 24.427, 24 October 2010. <<http://lists.digitalhumanities.org/pipermail/humanist/2010-October/001649.html>>.
- Humanist* 24.428, 25 October 2010. <<http://lists.digitalhumanities.org/pipermail/humanist/2010-October/001650.html>>.
- Robinson, Peter and Elizabeth Solopova. "Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue." *The Canterbury Tales Project: Occasional Papers*. Eds. Norman F. Blake and Peter Robinson. I vol. Oxford: Oxford University Computing Services, Office for Humanities Communication, 1993. 19–52.

- Robinson, Peter. "What Text Really is Not, and Why Editors have to Learn to Swim." *Literary and Linguistic Computing*, 24 (2009): 41–52.
Online: <<http://llc.oxfordjournals.org/cgi/content/full/24/1/41>>.
- Sakai Project: *An Open Source Suite of Learning, Portfolio, Library and Project Tools*. Sakai Foundation, 2010. <<http://sakaiproject.org/>>.
- Scriptorium. *Medieval and Early Modern Manuscripts Online*. Cambridge: University of Cambridge, 2006–2009. <<http://scriptorium.english.cam.ac.uk/>>.
- Simpson, John. "Using the OED as a Learning / Research Tool in Universities." *Digital Humanities 2006: Conference Abstracts*. Eds. Chengan Sun, Sabrina Menasri and Jérémy Ventura. Paris: Centre Cultures Anglophones et Technologies de l'Information, 2006. 290.
- Stokes, Peter A. "Palaeography and the 'Virtual Library'." *Digitizing Medieval and Early Modern Material Culture*. Eds. Brent Nelson and Melissa Terras. Arizona Center for Medieval and Renaissance Studies, forthcoming.
- Terras, Melissa. "The Digital Classicist: Disciplinary Focus and Interdisciplinary Vision." *Digital Research in the Study of Classical Antiquity*. Eds. Gabriel Bodard and Simon Mahony. Farnham: Ashgate, 2010 (a). 171–89.
- Terras, Melissa. "Present, Not Voting: Digital Humanities in the Panopticon." Plenary Lecture, DH2010, London, 10 July 2010 (b). ["Approximate" transcript online:] <<http://melissaterras.blogspot.com/2010/07/dh2010-plenary-present-not-voting.html>>.
- Tillotson, Diane. *Medieval Handwriting: History, Heritage and Data Source*. 2000–2010. <<http://www.medievalwriting.50megs.com/>>.
- Timemap. *Javascript Library to Help Use Google Maps with a Simile Timeline*. Google Code, 2010. <<http://code.google.com/p/timemap/>>.
- TNA: The National Archives. *Palaeography: Reading Old Handwriting 1500–1800, A Practical Online Tutorial*. <<http://www.nationalarchives.gov.uk/palaeography/>>.
- Tourelle, Jean-Claude. *Cours d'initiation à la paléographie Médiévale et Moderne*. 2nd ed. SIRA. <http://www-sira.u-bordeaux3.fr/moyen-age/cours_paleo/>.
- Treharne, Elaine M. "The Sensual Book and its Readers, 1000–1400: Keep your Wits About You." Leicester: unpublished lecture delivered 28 May, 2010.
- Turning the Pages*. London: British Library, 2010. <<http://www.bl.uk/onlinegallery/tpt/tpbooks.html>>.
- Twycross, Meg. "Teaching Palaeography on the Web." *Literary and Linguistic Computing*, 14:2 (1999): 257–284.
- Vogeler, Georg. "Einleitung: Der Computer und die Handschriften. Zwischen digitaler Reproduktion und maschinengestützter Forschung." *KPDZ* 1. xv–xxiv.
- [Vorholt, Hanna, Arianna Ciula, Peter A. Stokes and Elena Pierazzo.] *Manuscripts*. London: The Warburg Institute, 2010. <<http://warburg.sas.ac.uk/mnemosyne/word/manuscripts.html>>.
- Walsh, John. "Reaction to the Inadequacy of Embedded Markup." *Humanist*, 23 (2010): 776.2. <<http://lists.digitalhumanities.org/pipermail/humanist/2010-April/001199.html>>.

Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ?

Dominique Stutzmann

Résumé

Les initiatives TEI (Text Encoding Initiative) et MUFI (Medieval Unicode Font Initiative) multiplient les potentialités des transcriptions et invitent à interroger le processus même d'une opération pourtant commune dans les études médiévales. Transcrire, c'est décrire un texte et sa mise en forme graphique. Cela ouvre des voies aux études paléographiques, notamment par la description des diverses variantes morphologiques des lettres ou abréviations. Des recherches entreprises sur la production écrite de l'abbaye cistercienne de Fontenay démontrent le potentiel considérable des transcriptions dites « allographétiques » et des statistiques descriptives pour comprendre les évolutions de l'écriture médiévale. La fragmentation du paysage actuel de la recherche amène pourtant à souhaiter une harmonisation des pratiques, tant pour décrire les phénomènes paléographiques que pour structurer les données et formuler des ontologies.

Zusammenfassung

Elektronische Texte auf Basis der Text Encoding Initiative (TEI) und der Medieval Unicode Font Initiative (MUFI) erhöhen den Wert von Transkriptionen mittelalterlicher Handschriften so immens, dass dieser in der Mediävistik so gängige Prozess neu zu bewerten ist. Transkribieren heißt einen Text und dessen graphische Inszenierung beschreiben. Durch die Beschreibung der Buchstaben- und Abkürzungsvarianten werden der Erforschung der mittelalterlichen Schriften neue Wege geöffnet. Die Untersuchung der Buch- und Urkundenschrift der Zisterzienserabtei Fontenay zeigt, dass sogenannte allographetische Transkriptionen und die deskriptive Statistik einen beachtlichen Wert für das Verständnis der Schriftentwicklung besitzen. Die Vielfalt der heutigen Forschung in diesem Feld lässt eine Vereinheitlichung ihrer Praxis, sowohl für die Beschreibung der Formen als auch für die Datenstrukturierung und den Aufbau der notwendigen Ontologien, wünschen.

Abstract

Electronic texts employing the Text Encoding Initiative (TEI) and the Medieval Unicode Font Initiative (MUFI) provide added value for transcriptions of medieval manuscripts.

This is such a significant improvement that it requires rethinking the established process of transcribing. Transcribing is describing: with the encoding of variants for letters or abbreviations, palaeography can explore new horizons. A research on the scriptorium of Fontenay (O. Cist.) proves that graphetic transcriptions and statistical analysis improves our understanding of medieval scripts and their evolutions. Nonetheless, the field is diverse, and an elaboration of good practices for describing, structuring and organizing palaeographical data and relevant ontologies is urgently needed.

1. Introduction

Les humanités numériques et les éditions électroniques sont une chance pour les paléographes. La philologie étant de plus en plus attentive à la matérialité du texte, de nombreuses informations pouvant intéresser l'histoire de l'écriture sont consignées par les éditeurs de texte. Pourtant, malgré le temps passé et les nombreux projets dans le domaine philologique et paléographique¹, les transcriptions disponibles ayant un véritable intérêt pour une analyse paléographique sont rares. Il y a près de vingt ans, en 1993, P. Robinson et E. Solopova tiraient des conclusions critiques sur leur édition numérique de textes chaucériens (Robinson et Solopova 1993); les questions posées par ces deux auteurs demeurent d'actualité. L'amélioration des techniques ainsi que l'évolution des recommandations de la TEI (TEI Consortium) pour la structuration de l'information graphique ont profondément modifié les pratiques et les outils, mais trop de développements spécifiques empêchent encore une pleine collaboration.

Les enjeux scientifiques et industriels sont pourtant majeurs : du point de vue disciplinaire, il s'agit d'améliorer les classifications d'écritures et les identifications de mains, de créer de nouveaux critères de datation; du point de vue de l'ingénierie, l'appréhension des écritures anciennes et de leur variabilité pourrait bénéficier de l'existence de sources préparées et aboutir dans le futur à la reconnaissance optique des écritures manuscrites anciennes.

Le présent article essaie de définir les étapes nécessaires pour tirer bénéfice des possibilités d'analyse offertes par l'ordinateur : normaliser les transcriptions afin de favoriser les coopérations entre chercheurs et améliorer les outils d'analyse. Nous questionnerons dans un premier temps une dimension théorique de l'activité de transcription (que signifie transcrire ? à quel niveau transcrire ?), avant d'exposer la méthodologie et les résultats d'un projet spécifique de paléographie statistique. En confrontant différentes approches, nous proposerons un modèle et un cadre formalisé pour asseoir les pratiques futures.

¹ Une liste des corpus en langue française est présentée par C. Guillot *et al.* (2008).

2. L'image et la description du texte

Dans leur introduction, Robinson et Solopova abordaient la question théorique de la *signification de la transcription*, en affirmant que la transcription ne génère pas un substitut, mais constitue une suite d'actes de traduction d'un système sémiotique à un autre. La traduction, toujours incomplète et interprétative, est ainsi une pratique dont il s'agit d'interroger l'essence. Reposons la question : que signifie transcrire ?

2.1. Transcrire, c'est décrire

Il ne faut pas, selon nous, réfléchir à la transcription, même selon un encodage fin avec les éléments de la TEI, comme à la restitution ou à la « traduction » d'un texte, mais plutôt comme à sa description. Lors de la transcription, un premier document textuel donne naissance à un second ainsi qu'à de multiples descriptions : c'est ce que les sciences de l'information nomment « redocumentarisation » (Salaün 2008). Celle-ci intervient aussi bien lors du passage d'un « document » ou unité documentaire de l'analogique au numérique, avec la création de nouvelles métadonnées (descriptives, administratives ou techniques, en particulier sur la granularité, les relations entre documents et leur gestion) que lors de la constitution *a posteriori* d'un ensemble d'informations autonomes en « document ». Dans l'univers numérique, le texte même d'une ressource est indexable et exploitable : il double donc les indexations descriptives traditionnelles et devient sa propre métadonnée, ou, pour mieux dire, sa propre description à l'échelle 1:1.

Une structure de balisage telle que celle proposée par la TEI permet de décrire un « texte » en créant un nouveau document où l'opération descriptive va plus loin que la transcription, en explicitant des implicites du texte (par exemple, structure linguistique ou présence d'entités nommées, noms géographiques et anthroponymes). Cette possibilité nouvelle d'établir une description formalisée d'un texte à des échelles plus grandes que l'échelle 1, où le document descriptif dit davantage que l'original décrit, modifie profondément l'univers documentaire actuel et la conception des opérations documentaires : *transcrire* et *encoder* ne sont pas des opérations de traduction, mais des opérations descriptives et d'explicitation. Aussi de nouvelles questions se posent-elles : comment formaliser la description et ses différents niveaux ? Quel niveau de description choisir ? Faut-il étendre les règles internationales de description bibliographique (ISBD) pour rendre compte d'un texte et de sa description ?

2.2. Le texte est une image comme les autres

Robinson et Solopova proposaient quatre niveaux de transcriptions : « *graphic* » rendant toute la forme ; « *graphetic* » distinguant chaque type de chaque lettre ; « *graphemic* » préservant la suite de lettres ; « *regularized* » unifiant les suites de lettres attestées à une

forme normalisée. Si encoder, c'est décrire, l'on peut considérer sous un angle différent ces quatre niveaux allant des transcriptions « graphiques » aux éditions « régularisées » : les deux premiers décrivent l'image, le troisième décrit le texte-objet soumis aux accidents et le quatrième, le texte-idée. C'est la division fondamentale entre texte et image que nous retenons ici principalement. Dans un texte transmis par écrit, il y a toujours deux informations : le texte et l'image, le sens et la forme du signe².

La description de l'image, ou restitution formalisée des informations de formes, peut se diviser en deux niveaux. La transcription « graphique », tout d'abord, est définie comme restituant tous les caractères graphiques de l'original. Ainsi formulé, il s'agit d'une illusion, car nulle « transcription » ne peut donner accès à toutes les caractéristiques graphiques de l'original. En effet, la transcription porte sur le texte, alors que l'information décrite relève de l'image. Le document qui peut rendre compte de toutes les caractéristiques graphiques est la photographie, de sorte que la « transcription graphique » correspond à un ensemble descriptif constitué de métadonnées descriptives, d'une transcription du texte, de la photographie, avec association lettre à lettre des signes et de leur forme graphique (ensemble des coordonnées des points formant la lettre), voire les mesures effectuées par un logiciel d'analyse des formes. C'est, du reste, sur une « analyse graphique » plutôt que sur une « transcription graphique » que se fonde une part substantielle de la paléographie numérique, cherchant à reconnaître des mains et, partiellement, à améliorer la reconnaissance optique des caractères (Brink, Bulacu et Schomaker 2008 ; Bulacu et Schomaker 2007 ; Aussems et Brink 2009 ; Stokes 2009).

La transcription « graphétique » (*graphetic transcription*) ou, mieux, « allographétique » vise à donner accès à toutes les formes de chaque lettre ou signe³. C'est à elle que nous nous intéressons ici, car elle permet d'explorer des systèmes graphiques. Elle impose, pour ce faire, une réflexion sur les « types » et la réduction des variantes à des classes que l'on puisse désigner (cette nécessité confirme que la transcription allographétique n'est pas une transcription, mais une description). Il faut donc disposer d'un vocabulaire normalisé pour donner accès à certaines informations et asseoir sa pratique sur des règles qui permettent de prioriser les éléments à décrire, de hiérarchiser les informations selon les contextes et de traiter correctement les cas intermédiaires⁴.

² Cette distinction est une réalité largement problématisée dans les opérations de numérisation du patrimoine littéraire pour la constitution des bibliothèques numériques (André 2007).

³ La « graphétique » recouvrant l'étude des signes dans leur substance (formation, perception, lisibilité) et la « graphémique » celle des formes, l'étude « allographétique » est, malgré son nom, une branche de la graphémique (Catach 2001a ; Catach 2001b). C'est en raison de la polysémie des termes et des confusions possibles que nous retenons l'expression « transcription allographétique ».

⁴ Voir ci-dessous. Trois traitements sont en théorie possibles pour les cas intermédiaires : redoublement de l'information pour conserver toutes les interprétations possibles ; l'utilisation de valeurs prédéfinies permettant de trancher arbitrairement ; élaboration d'ontologies intégrant la double appartenance de valeurs mixtes.

La transcription allographétique pose ainsi des questions plus difficiles, car moins habituelles, que la *graphic transcription* qui réduit chaque forme à son sens dans un système alphabétique et où la tradition philologique a déjà clarifié les problèmes (Ecole nationale des Chartes 2001–2003). L'influence de ce que l'on peut appeler la « nouvelle philologie », attentive à la matérialité du texte (Cerquiglini 2000), oblige à remettre une partie de l'ouvrage sur le métier. La solution choisie pour les *Contes de Canterbury* est une description graphémique hybride qui consiste à transcrire les lettres, en ajoutant les abréviations composées d'une lettre avec un signe (comme les différentes formes de *p* barré) ou hors système alphabétique, ainsi que la ponctuation (Robinson et Solopova 1993). Ce système ne se justifie qu'avec la réduction du système abrégatif dans les langues vernaculaires (cf. Cottureau 2005 625–26). Si, dans les *Contes*, il n'y a que onze caractères abrégatifs codés hors lettres suscrites, ce type de transcription perd son sens dans des textes aux abréviations nombreuses et ambiguës : il accorde une place peut-être exagérée au système autonome de la ponctuation et pose implicitement l'hypothèse que la ponctuation et la capitalisation constituent une caractéristique plus structurante que les différentes formes des lettres, voire que les abréviations (approche commune dans les philologies non latines, cf. Lavrentiev 2007). Ce système hérite des préoccupations philologiques (savoir quelles parties du texte sont attestées ou restituées) et de la conception moderne de l'orthographe ; il étudie en conséquence le système graphique à partir du point de vue d'un alphabet unifié et décide *a priori* de ce qui est structurant et empêche par conséquent l'étude et la découverte de structures cachées.

La description allographétique est cependant possible : les réalisations actuelles semblent riches de promesses pour l'avenir. En exposant les résultats d'une analyse fondée sur de telles transcriptions, nous analyserons leur apport tout en insistant sur la nécessaire définition de « bonnes pratiques » et de normes communes qui assurent la pérennité des documents.

3. Un essai d'analyse de système graphique : corpus et méthode

Dans les paragraphes qui suivent, nous décrivons un projet et sa méthodologie, afin d'explicitier nos réflexions sur la structuration des données et réévaluer *a posteriori* la valeur heuristique de l'encodage allographétique utilisé. Trois exemples illustreront l'intérêt d'une analyse statistique des écritures sur la base d'un encodage limité ; ils permettront également de souligner les besoins actuels de normalisation et de tracer des pistes pour favoriser les enquêtes futures.

L'analyse allographétique et les résultats présentés ci-dessous ont été élaborés dans le cadre d'une thèse de doctorat où les différentes écritures attestées dans la production de l'abbaye bourguignonne de Fontenay aux ^{xii}^e et ^{xiii}^e siècles sont étudiées d'un point de vue statistique (Stutzmann 2009a). L'un des objectifs était d'identifier la production

par l'impétrant sédimentée dans le chartrier afin de confronter les caractéristiques paléographiques de la production pragmatique à celles de la production livresque.

3.1. Le corpus étudié

Le corpus initial était composé de 173 chartes originales antérieures à 1214 et l'analyse assistée par ordinateur a porté sur un sous-ensemble de 83 chartes et 40 livres manuscrits attribuables au scriptorium. Pour les livres, il a été procédé par sondage tandis que les chartes ont été encodées *in extenso*. Tous les écrits analysés sont en latin et en écriture gothique textuelle ou *textualis libraria* selon la typologie de Derolez (2003)⁵. L'exploration statistique a eu pour base un encodage suivant les recommandations de la TEI. Les résultats de l'analyse statistique sont particulièrement riches en regard des faibles moyens techniques mis en œuvre. Dans un système d'écriture très stable (écriture en langue latine, posée et dépourvue des phénomènes de cursivité permettant à l'individualité des scribes de s'exprimer), il apparaît tout de même possible d'étudier l'évolution des écritures ainsi que les divergences entre différentes mains, autant dans les taux d'utilisation des différentes formes que dans leur répartition et leur éventuelle régularisation.

L'étude de l'écriture dans une abbaye présente des caractéristiques particulières qui obligent à multiplier les approches et à élargir le champ d'investigation, puisqu'il ne faut pas se limiter à un auteur, mais en considérer plusieurs, dont le nombre et la qualité varient au fil du temps.

Au XII^e siècle et au début du XIII^e siècle, en Bourgogne du Nord, l'existence de plusieurs chancelleries peut être envisagée, dont un reflet pourrait se trouver dans le chartrier de l'abbaye cistercienne de Fontenay : celles des évêques d'Autun et de Langres tout d'abord, dont le diocèse s'étend sur une partie de la Bourgogne ducale (Richard 1954). La production documentaire des évêques d'Autun n'a pas encore été étudiée, tandis que celle des évêques de Langres fait l'objet des travaux d'Hubert Flammarion (cf. Flammarion 1982 ; Flammarion 2004). La troisième chancellerie dont on attend que le chartrier de Fontenay porte la trace est celle des ducs de Bourgogne où un bureau d'écritures s'organise sous Hugues III (1162–1192), même si les actes émanant des ducs ne se distinguent guère de ceux des barons bourguignons, et où une chancellerie se met réellement en place sous Eudes III avant 1218, avec des officiers et un formulaire imité de celui des rois de France (Richard 1984 381–84).

⁵ La classification élaborée pour les écritures minuscules gothiques par G. Lieftinck et augmentée par P. Gumbert et A. Derolez se fonde sur les morphologies de la lettre *a* (à simple ome ou à crosse), des hastes (bouclées ou non) et des lettres *f* et *s* long (sur la ligne ou plongeant) pour proposer des distinctions de types : *textualis* avec *a* à double panse, hastes non bouclées et *f* sur la ligne ; *cursiva* à l'opposé et *hybrida*, une *cursiva* sans boucle.

Ces trois autorités, disposant peut-être déjà de chancellerie, se retrouvent effectivement dans le chartrier, en compagnie de quelques autres⁶ :

- 81 actes, soit 46,8%, donnés sous le nom des évêques d'Autun ou de Langres, seuls ou en compagnie d'autres évêques ou abbés (53 pour l'évêque d'Autun et 30 pour celui de Langres, avec deux actes donnés en commun),
- 17 actes donnés par les ducs de Bourgogne (9,8%)⁷,
- 13 actes donnés par les archiprêtres de Touillon (7,5%)⁸,
- 7 actes donnés par les abbés de Fontenay (4%),
- 4 actes donnés par les archidiaques de Flavigny⁹,
- 3 actes donnés par les abbés de Flavigny¹⁰.

La répartition des autorités en diachronie n'est pas du tout homogène. Si les premiers actes, peu nombreux, sont placés sous le nom d'abbés, ceux rédigés entre 1150 et 1189 sont très majoritairement au nom d'évêques. À partir de 1190, l'importance numérique des autres auteurs croît fortement : les ducs de Bourgogne et l'abbé de Fontenay lui-même apparaissent de plus en plus fréquemment, ainsi que l'archiprêtre de Touillon, mais uniquement dans la décennie 1189–1199.

Si les autorités épiscopales et ducale sont bien attestées par le chartrier de Fontenay, leurs chartes ne forment cependant pas des ensembles cohérents et nulle trace de chancellerie constituée ne se reflète dans ce miroir. En revanche, plusieurs groupes d'actes se distinguent, qui dépassent les limites posées par le critère d'auteur. Une étude diplomatique et paléographique que nous ne reprendrons pas ici permet d'acquérir la conviction que leur élaboration est intervenue au sein de l'abbaye (Stutzmann 2009a 164–444). C'est sur les actes les plus anciens et ceux du scriptorium que se fondent les résultats présentés ci-dessous.

3.2. Méthodologie : encodage allographétique

En parallèle à l'étude morphologique, l'enquête paléographique a été menée sur le système graphique des scribes. La base en est un encodage allographétique respectant

⁶ La structure du fonds est similaire si l'on inclut les actes copiés, exception faite des bulles pontificales dont il ne subsiste qu'un original pour 17 actes connus par des copies.

⁷ Soit dix-huit actes si l'on compte l'un donné en commun avec les évêques de Lyon, Autun et Langres.

⁸ Le premier acte est donné en 1189 et les suivants entre 1194 et 1199, mais aucun entre 1200 et 1213. L'un de ces actes est donné sous les noms conjoints de l'archiprêtre de Touillon et l'abbé de Fontenay (ADCO 15 H 199/3).

⁹ Tous dans les quatre dernières années de notre période d'étude (un en 1210 et trois en 1213).

¹⁰ Actes répartis sur toute notre période (le premier, ADCO 15H257/ 1 datant de 1126–1149, le second 15 H 130 / 4 de 1180 et le dernier ADCO 15 H 58 / 1 de 1202); le second étant toutefois coémiss par l'évêque d'Autun.

les directives de la TEI-P5, effectué avec le logiciel Oxygen¹¹. L'encodage s'est fait sans souci de révéler des individus, mais avec l'objectif de voir des évolutions collectives. C'est un encodage de type générique qui a été choisi et ne portant que sur des phénomènes graphiques très largement répandus et observables y compris dans des textes courts.

Tous les actes ont été analysés à partir d'un seul fichier XML (chaque acte ou feuillet de manuscrit est encodé dans un élément <div>). L'analyse a été effectuée grâce à une transformation XSLT qui établit la liste des mots abrégés, calcule la largeur moyenne des lettres et décompte le nombre d'occurrences des allographes et des abréviations¹².

A partir des résultats chiffrés de la transformation XSLT, des analyses statistiques, essentiellement en composante principale, ont été réalisées avec logiciel R (Gentleman, Ihaka et R Development Core Team 2007).

Il faut bien noter ici que, comme toutes les études nécessitant un encodage des « caractères » d'une « population », celle-ci a opéré des choix qui ne sont pas neutres et la situent dans un champ où les tensions sont multiples : latin/vernaculaire ; écriture formelle/cursivité ; diplomatique/livresque ; individuel/collectif ; précision/généricité ; phénomènes ordinaires et fréquents/extraordinaires et rares.

En traitant de textes latins du XII^e siècle, l'alphabet présent est réduit : il y a peu de signes diacritiques (hormis la cédille de *e*) ; l'écriture ne présente pas de traits de cursivité dont le signalement modifierait substantiellement le travail de transcription et d'analyse : les ligatures canoniques *ct*, *et* et *st* apparaissent, mais sont presque seules, avec la fusion des oves contraposées qui apparaît sporadiquement. Les divergences morphologiques entre écritures de la pratique et écritures livresques sont en apparence minimales et ne peuvent pas être encodées directement avec MUFI (balancement des hastes et hampes).

La précision de l'encodage a été calibrée et mise en relation avec l'objectif. Les choix faits lors de cet encodage sont les suivants¹³ :

¹¹ Au moment de commencer cet encodage, la troisième version de MUFI (2009) n'avait pas encore paru, de sorte que certaines solutions d'encodage des signes abrégatifs sont personnelles.

¹² Les deux premières versions de cette transformation ont été programmées par Florence Clavaud, de l'École nationale des Chartes, que je tiens à remercier ici.

¹³ Le choix des encodages est justifié plus à plein dans le travail d'origine. C'est un choix qui s'insère dans une tradition paléographique ancienne. Dès les premiers traités diplomatiques, la forme des lettres comme critère de datation est expliquée, y compris pour les lettres *d* et *s*, mais les descriptions sont encore très sommaires (cf. Tassin et Toustain 1750–1765 : II, 167–73, étude de « *d* » avec une longue note sur la domination de la forme onciale à partir du milieu du XII^e siècle, et p. 260–72, étude de « *s* », en particulier à la p. 65 sur l'emploi des différentes formes dans les manuscrits). La première étude sur la forme du *d* est celle de Wilhelm Meyer qui, outre ses deux « lois » portant sur le *r* courbe après *o* et l'assimilation des oves contraires, constate une loi inconstante sur l'emploi des *d* : *d* droit devant lettres verticales *i*, *u*, *n*, (*m* et *r*) et *d* courbe devant lettre à oves *a*, *e* et *o* (cf. Meyer 1897 17–19). Bischoff prend soin d'indiquer la présence de *d* onciaux dans l'écriture des gloses (1954 8). Dans le scriptorium de Cluny, tous les scribes identifiés dans la deuxième moitié du X^e siècle utilisent les deux formes, mais leur utilisation n'est pas étudiée de façon différenciée (Garand 1978). Dans la bibliographie ultérieure, Petrucci affirme que l'emploi

- distinction de trois formes de la lettre *d* : capitale (« D »), *d* droit (« d »), *d* oncial (« Ɫ »)
- distinction de trois formes de la lettre *s* : capitale (« S »), *s* rond (« s ») et *s* long (« f »)
- encodage de toutes les abréviations avec leur résolution (par exemple : suite de balises
`<choice><expan>anima</expan><abbr>ai~a</abbr></choice>` pour le mot « anima ».

Outre cet encodage allographétique, des caractéristiques externes du texte (fin de ligne, espace blanc pour assurer la justification) ainsi que des éléments extra-paléographiques (noms de personnes et de lieu, mots en langue vernaculaire) ont été également enregistrés afin de pouvoir étudier leur influence sur le comportement des scribes. Des phénomènes ont été encodés aussi qui n'ont finalement pas été retenus dans le périmètre de l'étude (e.g. : degré d'abréviation et influence des fins de ligne).

3.3. Résultats obtenus

Au-delà des conclusions sur les pratiques paléographiques du scriptorium de Fontenay, le principal résultat de l'étude est que l'encodage de caractères limités (allographes *d* et *s*) a suffi à mettre en évidence des groupes homogènes et des évolutions, ainsi qu'à ouvrir la réflexion sur la perception de l'écriture dans le contexte médiéval. Plus

du *s* courbe en fin de ligne se constate en France dès les années 1140, puis en Germanie vers 1150 et à la fin du siècle en Italie, avant d'être d'usage régulier dans la seconde moitié du XII^e siècle en fin de mot (Petrucci 1968 1121–25). C'est aussi comme cela que nous interprétons la mention « (final) » sous le dessin de *s* courbe qui apparaît sur certains relevés de lettres de Gasparri (1973 28, 30, 38–39 etc). L'étude est poursuivie dans le contexte italien avec une approche systémique (Zamponi 1989 326–27), ainsi que dans le domaine allemand (Heinemeyer 1982 32–34 et 42–44). Dans le contexte espagnol, des exemples anciens de *d* oncial en finale précèdent la progression à partir de l'initiale et parachevée à la mi-XIII^e siècle. L'emploi des formes du *s* semble, plus que tout autre, répondre à la liberté des copistes et ne pas suivre d'évolution linéaire. Les autres allographes étudiés par Torrens sont *r* droit et rond, *i* et *j*, *u* et *v*, et *z* et *ç* (Millares Carlo et Ruiz Asencio 1983 I, 111, 85 et 94 ; Torrens 1995 355–59 pour *d*, 60–62 pour *s*). Si les planches sont représentatives, c'est la même évolution sur le plan local que nous constatons chez G. Nicolaj (1987 pl. XV) : apparition de *s* courbes dans la minuscule notariale qui en est dépourvue, en fin de mot et seulement après 1170. L'étude sur Gerhoh de Reichersberg, dont le scriptorium semble plus évolué et dont l'attention se porte sur les signes de ponctuation, ne fait pas le point complet sur la forme onciale de *d*, pourtant évoquée, (cf. Frioli 1999 207). A Brescia, la forme onciale, rare au début du siècle devient majoritaire aux alentours de 1150 pour s'évanouir ensuite et disparaître presque complètement dans les manuscrits liturgiques (Pantarotto 2005 5, note 25). L'approche morphologique est renouvelée par l'analyse des formes rondes comme élément d'un système graphique permettant de faciliter la lecture et d'assurer la compréhension en *lecture globale* (Frioli 2000 22–23). Il n'y a cependant pas d'évaluation statistique et l'interprétation pour la lettre *d* est très problématique : D. Frioli veut que le *d* oncial, après avoir marqué la fin d'une préposition (*ad*, *apud*) ou d'une forme pronominale (*quid*, *quod*), vienne à signaler le début ou la fin d'une syllabe, surtout si le mot est composé. Or la lettre *d* n'est, dans le système linguistique latin, jamais au milieu d'une syllabe. On en déduit donc qu'elle constate l'accroissement de la proportion de formes onciales.

concrètement, voici trois exemples choisis parmi les résultats positifs que cette méthode a permis d'obtenir.

Exemple 1

Le premier cas étudié est celui de la forme onciale de *d* dans un groupe d'actes écrits par le scribe principal des années 1150–1170. L'emploi des allographes de *d* est d'une nature dont ni la logique ni l'évolution n'apparaissent d'évidence. L'examen de tous les critères susceptibles d'influencer le scribe est trop complexe : il faut étudier la position dans le mot (initiale, médiane ou finale), les lettres précédentes et suivantes, ainsi que la présence de signes abrégatifs sur une lettre précédente ou suivante, voire sur la lettre *d* elle-même.

Les tableaux de chiffres rassemblant les cotes et les pourcentages selon chaque critère sont difficilement lisibles, d'autant qu'il faut pouvoir, dans chaque cas, avoir le nombre d'occurrences pour évaluer la représentativité d'un pourcentage. La table 1 indique les pourcentages d'emploi de la forme onciale selon la position dans le mot, sans le nombre d'occurrences, dans un tableau où les actes, majoritairement non datés, sont ordonnés selon leur *terminus ante quem*.

La lecture de ce tableau ne permet pas de tirer des conclusions immédiates. Il en est de même pour les pourcentages de formes onciales de *d* devant les voyelles *a*, *e*, *i*, *o* et *u*. Une analyse en composante principale, en revanche, fait apparaître une cohérence qui dépasse celle des séries linéaires. La figure 1 ci-dessous, qui est une représentation graphique de l'analyse en composante principale, met en évidence une évolution du système graphique au sein du scriptorium de Fontenay.

Les chartes originales sont dispersées selon plusieurs axes en fonction des pourcentages de formes onciales de *d* après les voyelles *a*, *e*, *i* et *u*¹⁴. Deux groupes apparaissent clairement et, dans la projection graphique, toutes les chartes antérieures à 1169 se retrouvent dans la partie gauche, tandis que celles postérieures se retrouvent à la droite. Cette distinction est extrêmement nette, au point qu'elle semble pouvoir être un critère pour proposer une datation de deux actes datables dans un intervalle assez long allant de 1163 à 1179 : d'un côté l'acte 15H 203/1 qui situe dans la partie gauche du graphique est plus vraisemblablement des années 1163–1169 ; de l'autre, l'acte 15H 249/1 se retrouve dans la partie droite du graphique, ce qui incite à le dater d'après 1169 sur des critères paléographiques, qui viennent ici corroborer un indice textuel ténu, puisqu'un témoin de l'acte n'est attesté que pour la période après 1171. Cette évolution concorde également avec d'autres évolutions de l'écriture dans le scriptorium (par exemple, multiplication de l'emploi de la forme ronde de *s* à partir d'environ 1170).

Cette mise en évidence est cruciale si l'on compare les chiffres bruts aux résultats de l'analyse en composante principale et à leur représentation graphique. L'analyse en

¹⁴ Le faible nombre d'occurrences de la séquence « do » empêche de tenir compte de l'influence de cette voyelle sur les formes de *d*.

Cote aux Archives dép. de Côte d'Or	Pourcentage de ð initial	Pourcentage de ð médian	Pourcentage de ð final
15 H 199 / 2 (s.d. [1154–1162])	14,3	23,1	100
15 H 156 / 2 (s.d. [1154–1162])	9,1	25	66,7
15 H 163 / 1 (s.d. [1154–1162])	25	46,2	100
15 H 163 / 3 (s.d. [1154–1162])	25	22,2	100
15 H 190 / 7 (s.d. [1154–1162])	18,5	37,2	72,7
15 H 193 / 2 (s.d. [1154–1162])	22,2	8,3	66,7
15 H 243 / 2 (daté 1162)	11,1	47,1	73,1
15 H 190 / 2 (s.d. [1163–1169])	25	37,5	66,7
15 H 190 / 4 (s.d. [1163–1169])	37,5	11,1	50
15 H 357 / 2 (daté 1169)	25	40	64,3
15 H 148 / 1 (s.d. [1148–1170])	15,4	17,6	66,7
15 H 163 / 2 (s.d. [1148–1170])	15,4	6,7	60
15 H 199 / 1 (s.d. [1162–1170])	6,3	0	66,7
15 H 249 / 2 (daté 1171)	51,4	61,4	87,5
15 H 249 / 3 (daté 1178)	39,1	46,4	91,7
15 H 193 / 1 (s.d. [1163–1179])	14,3	0	100
15 H 203 / 1 (s.d. [1163–1179])	29,4	20	68,8
15 H 249 / 1 (s.d. [1163–1179])	26,3	40	90
15 H 203 / 2 (s.d. [1171–1189])	40	52,6	100

TABLE 1. Pourcentage des formes onciales de la lettre *d* selon la position dans le mot.

composante principale offre un nouveau point de vue sur l'écriture et révèle l'influence du contexte (ici, la voyelle subséquente); elle permet d'affiner l'étude statistique et de rapprocher des écritures dont les pourcentages moyens divergent fortement. Ainsi, les deux actes 15 H 163/1 et 15 H 163/3, très proches l'un de l'autre sur les critères tant internes qu'externes, se retrouvent très proches sur le graphique alors que leurs pourcentages de *d* onciaux en position médiane varient du simple au double.

Ce premier exemple montre qu'une analyse statistique du matériel paléographique est possible et pertinente, y compris pour obtenir de nouveaux indices de datation. Il montre également que l'enquête statistique permet d'approcher de façon neuve des mécanismes d'écriture invisibles ou impossibles à étudier autrement, tels que l'influence des lettres subséquentes sur la graphie d'une lettre précédente.

Exemple 2

Le deuxième exemple choisi est l'évolution de l'emploi de la forme onciale de *d* entre 1150 et 1189, période durant laquelle le scribe principal des années 1150–1180 se voit doter de deux collègues actifs dans les années 1170–1189. Le procédé est le même : une

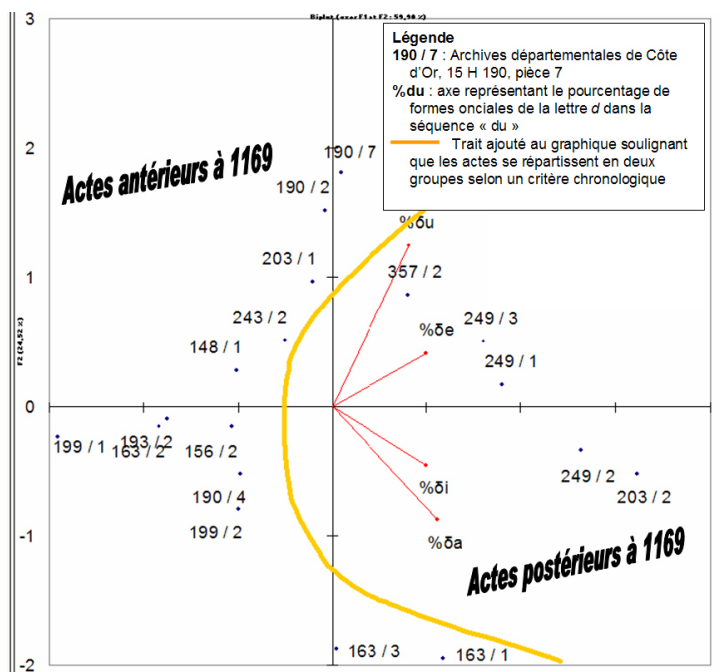


FIGURE 1. Analyse en composante principale mettant en lumière l'évolution du système graphique.

analyse en composante principale est effectuée, en tenant compte de la voyelle qui suit immédiatement la lettre *d*, et les résultats de cette analyse sont représentés en deux dimensions.

La frontière déjà observée vers 1170 se retrouve, mais surtout l'ensemble permet de constater une évolution plus générale.

Le premier graphique ci-dessous montre que ces deux scribes (ici appelés « 2a » et « 2b »), plus jeunes, ont des habitudes graphiques globalement distinctes de celles de leur aîné, mais qui ne se distinguent pas entre elles. Dans ce cas-ci, ce n'est pas l'évolution dans la production d'un même scribe que l'on fait apparaître, mais des évolutions générationnelles et partagées par plusieurs individus.

L'examen étendu à la production des années 1190–1215 fait, lui aussi, apparaître des divergences entre les pratiques de différents scribes, mais les séparations sont plus progressives et moins tranchées, alors que du point de vue morphologique, des caractéristiques distinctives permettent d'isoler les écritures tardives. Nous pouvons ainsi conclure de l'analyse statistique des phénomènes paléographiques que des

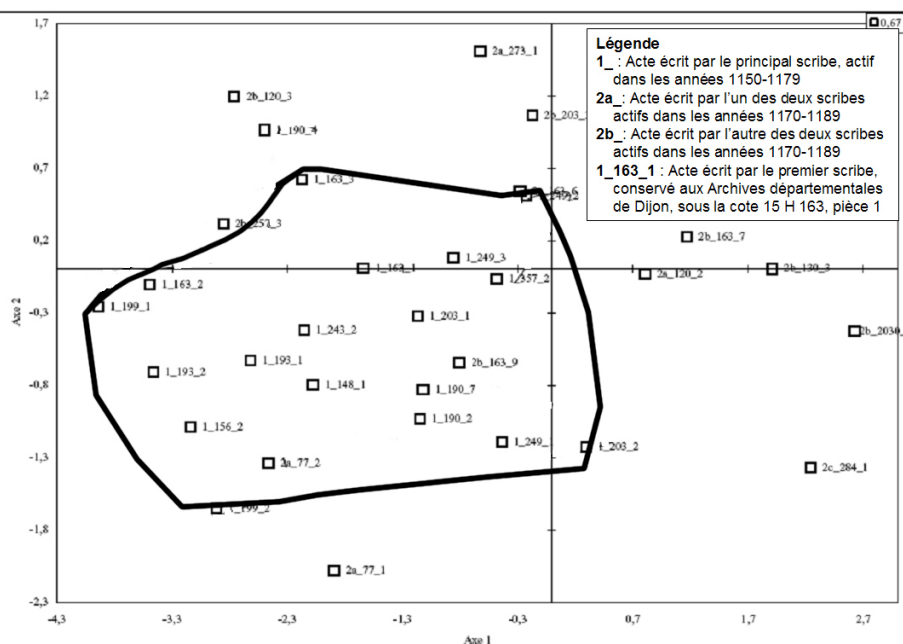


FIGURE 2. Analyse en composante principale montrant la différence entre le système graphique d'un scribe d'une génération antérieure et celui partagé par deux scribes d'une nouvelle génération.

glissements générationnels interviennent, affectant séparément la morphologie des lettres et le système graphique dans son ensemble.

Exemple 3

Le troisième exemple montre la liaison possible entre des caractéristiques d'écriture et des réalités extra-paléographiques, en particulier décoratives. En effet, dans les livres manuscrits de l'abbaye de Fontenay, plusieurs relèvent du style « monochrome », déjà étudié en détail pour Cîteaux par Y. Załuska (1989).

Ce style monochrome, qui est une création cistercienne, impose que les initiales soient d'une seule couleur. Il n'interdit pas l'emploi de plusieurs couleurs sur une même page et n'entraîne pas d'économie de pigment : au contraire, il se développe une technique graphique qui joue avec le blanc du parchemin, autorise de somptueuses initiales comme dans la Bible de saint Bernard et permet même un jeu multicolore d'initiales enclavées.

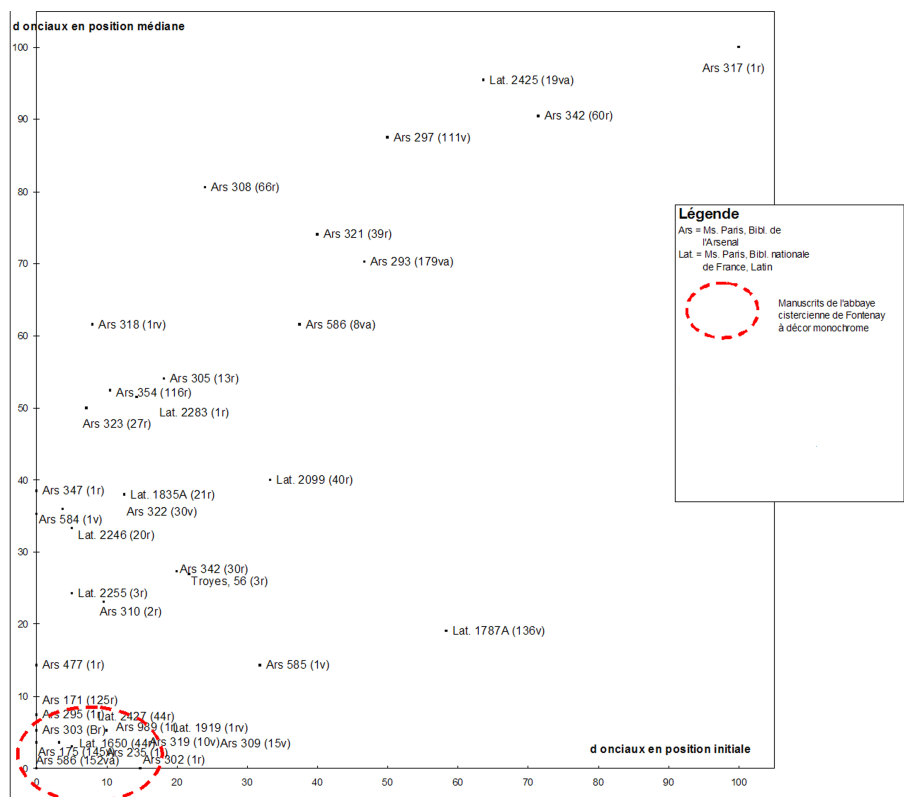


FIGURE 3. Régularisation de l'écriture dans les manuscrits à décor monochrome (spécialisation de la forme onciale de *d* pour la fin de mot).

A Fontenay, l'analyse paléographique et codicologique démontre que ces manuscrits partagent de nombreuses spécificités outre leur décoration. En particulier l'emploi des formes droite et onciale de la lettre *d* montre une spécialisation des formes de cette lettre *d* dans les manuscrits à décor monochrome, où la forme onciale est exclue des positions initiales et médianes et strictement réservée à la finale.

L'utilisation d'une forme spécifique en position finale permet d'améliorer la lisibilité d'un texte où les espaces entre les mots sont irréguliers en soulignant, par la morphologie, un phénomène linguistique (Frioli 2000). Cependant il ne faut pas réduire ce processus de spécialisation à une seule recherche de lisibilité : en effet, si celle-ci était le seul

enjeu, elle eût pu être maintenue hors des modes décoratives. Or, non seulement, ce phénomène est transitoire, mais il intervient précisément dans les manuscrits qui se distinguent par leur décor soumis à la norme nouvelle de la monochromie ainsi que par d'autres caractéristiques matérielles, tels largeur des marges ou grandeur de l'unité de réglure (Stutzmann 2009a). Une telle règle ou régularisation graphique sur les formes de la lettre *d* est en outre totalement inconnue à la production documentaire. Cela vient, à notre sens, confirmer qu'il s'agit de l'effet d'une mise en ordre délibérée des réalités paléographiques et pratiques décoratives dans les livres manuscrits. De même que l'interdiction de la polychromie peut être interprétée comme une conséquence d'ordre esthétique de la théorie des sens et du salut chez Bernard de Clairvaux (Stutzmann 2009b), il apparaît que la régularisation de l'écriture, qui est soumise à une contrainte nouvelle ou à une règle, correspond à une volonté de maîtrise des sens et à une réflexion sur l'ascèse.

3.4. Conclusions méthodologiques

Les trois exemples allégués ci-dessus, exploitant des données simples et des ensembles à faible population, appellent des conclusions méthodologiques. Des analyses statistiques, notamment factorielles, permettent d'exploiter des informations dont la pertinence n'est pas prédictible et de mettre au jour des cohérences de pratiques scripturales malgré la forte divergence des données. L'encodage allographétique permet d'interroger sous un angle supplémentaire les données paléographiques et de découvrir des évolutions substantielles de l'écriture.

L'analyse statistique, fût-ce d'un unique caractère, constitue un critère déterminant pour l'enquête paléographique, car le processus d'évolution des écritures ne porte pas seulement sur la morphologie des lettres, mais aussi sur l'agencement des allographes. C'est le système graphique en son entier qui évolue, y compris au sein d'une production limitée dans le temps et l'espace.

4. L'analyse des écritures : normaliser pour coopérer

La conclusion positive issue du processus analytique précédemment exposé donne naissance à de nouvelles interrogations. Est-il envisageable de traiter de façon similaire des corpus à la fois plus vastes et moins homogènes ? Quel rôle attribuer à la machine dans l'analyse des écritures ? Dans le cas présenté ci-dessus, l'analyse statistique permet de souligner les phénomènes de cohérence ou de divergence au sein d'un ensemble qui se présente comme homogène et peu variant. Les conclusions plus profondes s'obtiennent en croisant les informations sur les pratiques d'écriture, les réalités codicologiques, l'enluminure et la connaissance de l'univers intellectuel médiéval.

Pour étendre l'enquête et affiner la méthodologie, il faut surmonter le défi de la masse des données et exploiter de façon adaptée et pertinente des données hétérogènes. En conséquence se pose la question des pratiques d'encodage utilisées et des fonctionnalités requises pour les outils de traitement de l'information paléographique.

Encore rares actuellement, les données paléographiques s'accroissent en effet rapidement. A l'heure actuelle, chaque projet mené par des équipes aux objectifs différents élabore ses propres solutions spécifiques, même si TEI et MUFI offrent un canevas commun. Des oppositions tendent à disparaître, notamment celles consistant à désigner des phénomènes graphiques différents par un même descripteur : par exemple « s » soit pour *s* rond (solution la plus fréquente), soit pour *s* long, l'autre forme pouvant être déclarée contradictoirement sous la même entité « &s ; » (Uitti 1997 ; McGillivray 1994–2010). Il est en revanche naturel qu'un projet portant sur un corpus spécifique où un allographe est nettement dominant, voire d'utilisation exclusive, tende à transcrire celui-ci sous la forme équivalente de l'alphabet latin actuel, indépendamment de son inclusion dans l'histoire de l'écriture médiévale : ainsi l'on trouvera les caractères « d » et « f » pour transcrire une forme onciale bouclée ou un *s* plongeant sous la ligne, y compris dans des projets avec une granularité descriptive forte, allant jusqu'à avoir plusieurs types de *i* court (Hofmeister, Hofmeister-Winter et Thallinger 2009). La lourdeur de préparation des fichiers et la diversité des objectifs poursuivis exigent des solutions économiques et adaptables aux besoins de chaque équipe.

Or, dans notre première partie, nous avons évoqué la possibilité de décrire les informations à une échelle supérieure à 1. Si l'analyse et les transcriptions allographétiques portent, selon les projets, à la fois sur les écritures livresques et celles de la pratique et qu'elles couvrent plusieurs des siècles médiévaux, le nombre de documents concernés peut rapidement atteindre des proportions hors des capacités d'analyses individuelles, proportions où chaque signe devient une information à analyser et exploiter pour interpréter le système graphique d'un document. Le nombre en est potentiellement infini. Même si les informations paléographiques sont homogènes et nonobstant les difficultés d'analyse, l'on court le risque de ne pouvoir interpréter les similitudes entre écritures relevant de familles graphiques différentes. Des similitudes démontrées par l'analyse statistique seront une donnée supplémentaire à interpréter, et non une preuve démonstrative. Il est en effet difficile d'établir des arguments et des valeurs causales dans un domaine strictement paléographique et sans donnée extérieure pour appuyer l'interprétation.

L'amélioration des analyses exige ainsi un plus grand nombre de fichiers structurés avec des informations de nature allographétique et la possibilité de disposer de corpus de comparaison pour valider les hypothèses et résultats. La pluralité des choix d'encodage nécessite d'établir des passerelles, des normalisations ou des bonnes pratiques (valeur des caractères, structuration, granularité descriptive) et une formalisation des choix

précisant les pratiques adoptées, afin de pouvoir traiter de grandes masses de données partiellement hétérogènes.

4.1. Valeurs

Le premier niveau auquel on pense est naturellement celui des valeurs utilisées : quel caractère utiliser pour quelle forme graphique. Paradoxalement, ce sont les signes non-alphabétiques ou de l'alphabet non latin qui font l'objet de la plus grande attention : signes graphiques adventices, abréviations, capitales, après un essai infructueux avec *r* et *s* pour Robinson et Solopova (1993), lettres anglo-saxonnes pour Rumble (2004), abréviations, *s* et ligatures pour McGillivray (2005). L'étude minutieuse des pointages de *i* et de plusieurs autres caractéristiques a pu être menée sur des corpus restreints et constitue encore un cas particulier (Hofmeister, Hofmeister-Winter et Thallinger 2009). Certaines listes sont particulièrement longues (Uitti 1997 ; McGillivray 1994–2010). Les projets portant sur des corpus vernaculaires anglo-saxons, germaniques ou nordiques, dont les particularités graphiques imposent l'usage de signes graphiques hors de l'alphabet latin de 22 lettres ont déjà abouti à des réalisations fondatrices et bien documentées.

La question de la normalisation et de l'évolution d'Unicode est parfois explicitement mentionnée, en particulier dans des projets liés à MUFI (Haugen 2008 ; Anderson et al. 2007 ; MUFI 2009). Cette initiative répond en effet à de nombreuses questions, mais pas encore toutes, et, parfois, engendre de nouvelles difficultés. Ainsi la forme plongeante de la lettre *s* : MUFI permet d'encoder une forme agrandie (code EEDF) ou la forme insulaire (code A785), mais pas la forme fréquente du *s* de la bâtarde bourguignonne. Cela pourrait être amélioré dans les prochaines versions à condition de considérer ces allographes comme des caractères et non comme des formes ; une réflexion approfondie et une enquête nouvelle doivent être menées pour définir l'allographie selon ce nouveau point de vue.

Outre la question des entités et valeurs à utiliser pour décrire les différentes morphologies d'une lettre, la principale question concernant les valeurs tient à la double possibilité d'utiliser un encodage spécifique ou des caractères génériques. Les lettres suscrites, notamment, donnent presque toutes occasion à variantes d'encodage : les voyelles et semi-voyelles (*a*, *e*, *i*, *o*, *u*, *w*) avec lettre suscrite ont un code particulier, alors qu'existent par ailleurs les lettres suscrites seules (par exemple 0363 pour *a* suscrit, autrement dit *combining latin small letter a*). Les ligatures, au contraire des lettres suscrites, sont systématiquement traitées par MUFI comme un unique caractère spécifique et ne donnent pas lieu à un codage générique, avec pour principaux inconvénients l'oubli inévitable de cas rares (par exemple, ligature ou plutôt lettres conjointes *nt* avec forme capitale de *n*) et la réduction du décompte du nombre de lettres — la solution choisie n'est pas indifférente pour des projets où les lettres sont décomptées

pour calculer la densité graphique ou le degré d'abréviation d'un texte (Cottureau 2005 ; Bozzolo et al. 1997)¹⁵.

Les artefacts de présentation tels qu'agrandissement ou étirement d'une lettre sont aussi traités comme caractères spécifiques, de sorte que un *a* carolin et un *a* carolin agrandi seront deux caractères distincts. Or, d'un point de vue paléographique, on peut considérer qu'il s'agit de la même morphologie. Pour traiter une première ligne d'un acte carolingien en lettres étirées ou un mot en lettres agrandies, deux possibilités contradictoires apparaissent : soit coder chaque caractère spécifiquement, soit marquer une chaîne de caractères avec un qualificatif de mise en forme (le cas échéant, la valeur de celui-ci devrait également être normalisée, par exemple « enlarged », « elongated », comme dans `<hi rend="enlarged">`), de façon à rendre possible les conversions vers les fontes qui nécessitent effectivement que chaque caractère ait un code.

Un cas limite du respect de la morphologie et de la nécessité du sens est soulevé par Robinson et Solopova (1993) : l'existence de formes identiques pour des lettres différentes, telles que *c*, *e*, *o*. Deux formes différentes de codage sont possibles :

- soit transcrire avec un caractère explicitant uniquement la forme (par exemple « *c* » pour un *e* non bouclé) et indiquer la forme normalisée (en TEI, `<c rend="c">e</c>` ou, plus complexe, `<choice><orig>c</orig><reg>e</reg></choice>`),
- soit transcrire avec un caractère spécifique, à déterminer, la forme de *e* qui ressemble à un « *c* ».

Dans ces deux solutions ne se pose plus seulement un problème de valeur de caractère, mais de structuration de l'information et de granularité descriptive.

4.2. Structuration

Les mises en formes ou les morphologies ambiguës ne sont pas les seules réalités paléographiques qui peuvent être rendues soit par des caractères spéciaux, soit par un balisage et une structuration alternative. Ce sont les abréviations qui posent le mieux le problème de la structuration de l'information. Certains projets y ont consacré une grande attention, mais la diversité est encore de mise (Lavrentiev 2002 ; Heiden, Guillot et Lavrentiev 2002–2008). Le débat n'est pas clos et les réflexions, menées principalement pour les langues vernaculaires, doivent encore être affinées (Mazziota 2008).

¹⁵ Pour l'analyse de l'espace écrit, ligatures et abréviations constituent des cas d'études distincts. La lettre suscrite n'est pas omise, mais ne modifie pas l'espace d'écriture occupé par la lettre qu'elle surplombe ; la présence d'une ou plusieurs lettres adscrites, au contraire, est à évaluer dans l'analyse de l'espace écrit, mais leur module réduit rend l'évaluation difficile ; les ligatures, enfin, posent aussi problème puisque la longueur du digramme est susceptible d'être modifiée, mais ne l'est pas toujours (les ligatures *st*, *fi* ou *ti* sont-elles plus courtes que les deux lettres séparées ?) et il est délicat de ne le tenir que pour un seul caractère.

La TEI-P5 prévoit un système complet pour déclarer une abréviation (<g>, <char>, <glyph>) et pour décrire le mot : soit <abbr> et <expan> pour le mot entier (abrégé ou avec abréviations résolues), soit <am> et <ex> pour chacune des abréviations. Ce système pose des problèmes de cohérence et d'exploitation. Certaines abréviations par contraction syllabique apparaissent construites sur un radical puis déclinées, et autorisent ainsi plusieurs encodages divergents : dans le mot *anima* abrégé *a*, *i* tilde, *a*, l'abréviation est-elle composée de *a*, *i* tilde, ou de l'ensemble du mot ? et auquel cas, peut-on encore utiliser <am> ou bien doit-on utiliser <abbr> ? et alors faut-il considérer que *anima* et *anime* sont le sujet de deux abréviations différentes, et quel lien établit-on avec *animus* ?

Dans le cas d'autres abréviations par contraction, si l'on fait le choix de mettre en évidence le radical, celui-ci prend lui-même les différents degrés : *frater* abrégé sur le radical *fr* se décline en *fratrum* marqué « frm » et tilde, ou encore *homo*, *-inis* avec les abréviations tildées *ho*, *hoie*, *hois*, *hoim* où les radicaux *ho* et *hoi* signifient respectivement *homo* et *homin-*, *homini* et *hominu-*. Un très grand nombre de mots courants sont l'objet des abréviations par contraction et suppriment un nombre variable de lettres (par exemple : *abbas*, *dominus*, *gratia*, *martir*, *nomen*, *noster*, *omnis*, *pater*, *peccatum*, *sanctus*, *seculum*, *spiritus*, *vester*, ou encore *ecclesia* aussi bien *eccla* ou *ecclia*). Ces abréviations, à la fois courantes et simples, obligent à donner systématiquement la résolution. Des mots peuvent connaître plusieurs abréviations apparentées : faut-il en considérer certaines comme par suspension et d'autres par sigle abréviatif (exemples : *presbiter*, abrégé soit *pbr* par contraction et *p* macron, *s*, *b* barré, *r* par sigles abréviatifs) ? D'autres abréviations par contraction, notamment avec lettre suscrite, sont ambiguës et exigent leur développement (ex. : *m* avec *o* suscrit pour *modo* et *monacho*) : évidemment les abréviations par suspension, mais aussi les signes abréviatifs les plus courants peuvent être ambigus (*d* barré pour *de*, *-dem* ou *-ud*; *p* barré pour *per*, *par* et *por*, etc.), tandis que le simple encodage par résolution n'est pas toujours suffisant pour savoir quelle abréviation a été utilisée (« *que* » peut valoir pour « *q* », « *qz* », « *q* » avec *e* suscrit, ou « *q* » barré ; « *bus* » pour « *b* », « *bz* » ou « *b*⁹ » ; « *com* » pour le neuf tironien, le *c* tildé ou le *c* retourné ; « *esse* » pour « *êê* » ou « *eê* » ; « *rum* » pour *r* rond barré et *r* tildé). L'étude des abréviations doit, par ailleurs, pouvoir tenir compte des alternances allographétiques (par exemple *dictus* abrégé *d*, *c*, tilde avec *d* droit ou oncial).

Les solutions proposées par Mazziota pour les abréviations sont complexes et s'opposent aux tendances de MUFI et de l'Unicode (par exemple il ne faudrait pas considérer *p* barré comme un signe abréviatif autonome ou « logogramme », mais le décomposer en lettre *p* et signe distinctif ou « cénégramme »). Elles s'enferment aussi dans des barrières de modélisation qui nous semblent contraire avec la perception globale des mots, en refusant le traitement des abréviations par contraction et en imaginant des « périgrammes » (sous-type des cénégrammes) à « portée discontinue » (Mazziota 2008 : §46). Ces propositions ont pourtant des avantages : elles permettent de

rendre compte du déplacement du signe abrégatif ou de superpositions et de décrire très finement l'abréviation tout en unissant structurellement les lettres restituées au signe abrégatif.

A des fins de coopération, il serait souhaitable de pouvoir unifier la structuration, ou, pour le moins, de formaliser les choix d'encodage, car celui-ci dépend en effet de la compréhension du système médiéval et l'unanimité demeure incertaine. Aussi les pratiques adoptées doivent-elles être documentées (en TEI dans l'élément <encodingDesc>) et formalisées.

4.3. Granularité descriptive et explicitation du référentiel utilisé

Au-delà du choix des valeurs utilisées et de la structuration de l'information paléographique, un choix encore plus fondamental doit être précisé : la granularité descriptive, ou degré de précision de l'encodage. Celle-ci porte d'une part sur la globalité du système d'écriture, d'autre part sur les morphologies des lettres.

Pour analyser de grandes masses de données, deux logiques, en apparence opposées, coexistent : d'une part, celle de « l'exploration des données » et des moteurs de recherche sémantiques, approche microscopique cherchant à analyser les informations dans leur plus grand détail afin d'établir *a posteriori* une interprétation hors d'un modèle probabiliste et hors des modèles d'interprétation *a priori* ; les statistiques descriptives et analyses factorielles se rattachent à cette famille d'analyse. D'autre part, la logique descriptive catalographique, perspective de télescope, qui cherche à rendre accessible et visible des ensembles vus dans leur globalité et ne s'appréhende que par des points d'accès prédéfinis et un vocabulaire contrôlé, éventuellement coordonné et hiérarchisé, formant une ontologie.

Dans le domaine documentaire, plusieurs travaux montrent que les meilleurs résultats sont obtenus en confrontant les référentiels structurés, élaborés par analyse intellectuelle, et la fouille de données, fondée largement sur l'analyse statistique et portant sur des documents non structurés (Mane 2010 17–27 ; Beneventano et Bergamaschi 2006 ; Criado Fernández et Martínez-Tomás 2009). Dans le domaine purement paléographique, c'est également la double approche qui semble être la plus prometteuse (Muzerelle 2009 6–9 ; Hofmeister, Hofmeister-Winter et Thallinger 2009) : apporter une description télescopique globale (« textualis », « hybrida », etc.) et examiner les caractéristiques comme au microscope (*s* rond, *i* pointés, hastes bouclées, orientation des traits etc.).

Pour que chaque corpus puisse servir de point de comparaison aux autres, de façon neutre et sans préjuger des recherches de chaque projet, il est indispensable de pouvoir déclarer non seulement les caractères qui sont encodés, mais aussi ceux pour lesquels aucune supposition ne peut être faite. A cette fin, l'utilisation de typologies d'écritures, autrement dit d'ontologies, apparaît la meilleure solution. En effet, pour comparer des données en très grand nombre et structurées selon des codes hétérogènes, deux solutions

sont envisageables en théorie : convertir et réduire chacun des documents produits pour le soumettre à une norme unique, ou élaborer un système qui permette au document de déclarer selon quel système est réalisé l'encodage. La conversion de données signifie souvent la réduction au plus petit commun dénominateur et engendre une grande perte d'informations. C'est donc plutôt la seconde solution qu'il faut encourager et favoriser : définir plusieurs niveaux descriptifs, ayant chacun ses caractéristiques

L'emploi d'une classification telle que celle de Lieftinck-Gumbert-Derolez (Derolez 2003) permet de préciser la pratique d'encodage au niveau global du document et évite de préciser systématiquement la forme présente dans l'original. Ainsi, l'emploi des dénominations « *textualis* » ou « *hybrida* » permet d'interpréter les caractères *b*, *d*, *l*, *h* comme désignant des morphologies dépourvues de boucle (inversement pour la *cursiva*), indépendamment de leur encodage au sein du document traité et indépendamment de l'acceptation faite par ailleurs du système de classification de l'écriture. En revanche ces mêmes dénominations distinguent que partiellement les formes de *s* si la différence n'est pas faite entre *s* longs et *s* ronds. La précision du système doit être insérée soit au niveau du document, soit à l'intérieur même de la transcription allographétique. Imaginons, en effet, le cas des manuscrits dont les débuts de chapitres sont en *textualis libraria* et le corps en *cursiva* : sans une granularité et une explication suffisante, les caractères « *a* » ou « *f* » ne pourront être simplement pas être exploités si l'on veut distinguer les allographes *a* à simple boucle et *s* plongeant.

La granularité descriptive joue également un rôle dans l'encodage allographétique pour chaque lettre à formes multiples. Robinson et Solopova justifient le rejet d'une transcription allographétique par la multiplicité des morphologies et renvoient aux huit formes de *s* de Benskin (1990), en insistant sur le développement pyramidal des graphèmes : « *It is assumed by both McIntosh and Benskin that the relationship of graphemes to graphetes is hierarchical : so many graphetes of s ; so many sub-types of each graphete ; even sub-sub-types, and so on.* » (Robinson et Solopova 1993).

C'est bien là, pourtant, que réside une solution : adopter une hiérarchie et encoder, au besoin, à niveau élevé, en acceptant la part d'arbitraire qui réside dans la nature même de la description par encodage. Pour des comparaisons à large échelle visant à comprendre les évolutions du système graphique, des distinctions objectives simples pourraient sans doute suffire — et seule la mise en œuvre réalisera la preuve de concept —, même si la hiérarchie devrait, en bonne théorie, être construite par le bas, en repérant toutes les formes et leurs contextes d'apparition pour pouvoir établir des types. Les travaux d'Oeser sur les formes de la lettre *a* et des jambages montrent bien la signification que peuvent avoir des variations morphologiques (Oeser 2001–2002 ; Oeser 1994), mais n'invalident pas la distinction fondamentale entre *a* à simple ove et *a* à double panse. L'on peut, à l'envi, raffiner sa description ou la simplifier, indépendamment de l'absence de consensus parmi les paléographes sur ce qui fonde un type. Il s'agit là de choix

d'encodages et de granularités descriptives différentes : ceux-ci doivent être réfléchis et explicités de façon à préserver l'interopérabilité sémantique des données.

5. Les outils

Dans les paragraphes qui précèdent, nous avons décrit les besoins de normalisation et d'harmonisation concernant les valeurs, la structuration et la granularité, de façon à obtenir des fichiers de travail dont le contenu soit explicite et utilisable par un traitement automatisé d'informations nombreuses sans générer de bruit. L'accent y a été mis délibérément sur la structuration des fichiers contenant des informations paléographiques fines. Néanmoins la question des outils est tout sauf accessoire, aussi bien en amont, pour la constitution de bases de données paléographiques, qu'en aval, pour leur exploitation. Les outils génériques n'existent pas encore, qui permettraient de proposer une nouvelle perspective sur l'ensemble des écritures médiévales. Examinons tout de même les fonctionnalités souhaitables des logiciels d'aide à l'analyse paléographique.

5.1. Constituer des bases de données paléographiques

En amont, dans la phase d'enregistrement des informations paléographiques et dans la structuration même, tout le labeur ne doit pas être pris en charge par les chercheurs et la machine ne sert pas qu'aux calculs. L'équipe de *DAmalS* a exprimé à quel point l'encodage allographétique et la mise en relation du texte transcrit avec l'image correspondante sont aussi lourde qu'importante (Hofmeister, Hofmeister-Winter et Thallinger 2009 270, 276). La création d'un logiciel de prétraitement des textes manuscrits médiévaux serait très précieuse. En offrant une segmentation (par mots et caractères, ou par groupes de caractères liés), éventuellement des fonctions de reconnaissance des formes à l'intérieur d'une même page (pour repérer les mots répétés) ainsi qu'une interface encourageant la saisie des caractéristiques paléographiques pour chaque zone repérée, un tel outil permettrait non seulement de faciliter le travail des chercheurs, mais également d'assurer la normalisation et de diffuser les bonnes pratiques pour disposer de données interopérables, tant pour les valeurs et descripteurs utilisés, la structuration et l'explicitation du niveau d'encodage utilisé.

Lors de la réalisation d'une transcription allographétique, il serait *primo* souhaitable de disposer d'une interface présentant une lettre isolée, un unique mot ou une seule ligne à transcrire. Du point de vue logiciel, cela correspond aux problématiques de segmentation, c'est-à-dire de reconnaissance des unités graphiques et des espaces entre les lettres, mots et lignes. Les algorithmes de segmentation sont complexes (Gatos, Stamatopoulos et Louloudis 2010) ; ce sont des boucles récursives qui analysent la probabilité de chaque hypothèse de segmentation et les confrontent d'un côté à

des dictionnaires et de l'autre aux résultats de l'analyse des formes effectuée selon l'hypothèse considérée (Tzadok et Walach 2009). Des essais effectués à la Bibliothèque nationale de France montrent que la fonction de segmentation du logiciel FineReader (société ABBYY) obtient déjà des résultats corrects, même si elle n'est pas tout à fait mûre pour les manuscrits médiévaux en raison de l'impossibilité de confronter efficacement l'hypothèse de segmentation à un dictionnaire.

Secundo, des hypothèses de transcription pourraient être proposées. Dans des corpus suffisamment homogènes, une interaction entre les algorithmes de segmentation et des fonctionnalités d'apprentissage serait bénéfique. Des logiciels peuvent même déjà se vanter de résultats corrects d'apprentissage, tant pour les mots, ou « word-spotting », que pour les lettres (Tomasi et Tomasi 2009 ; Leydier, Duong et Ouji 2006–2009).

Tertio, face à une image ou à une proposition de transcription, le paléographe a besoin d'une interface adéquate, rendant aisée la tâche d'analyse paléographique. Cela signifie notamment que le chercheur doit disposer d'un dictionnaire des entités et des abréviations facilement utilisable et lié à l'éditeur XML.

Quarto, afin de préparer l'exploitation des données, la constitution d'une base de données paléographiques implique de conserver les informations reliant la transcription allographétique réalisée et l'image originelle.

La situation est insatisfaisante. D'un côté les outils d'annotations actuels permettent d'enregistrer l'information, mais sans offrir ni aide à la saisie, ni structuration suffisante, de l'autre les outils de segmentation ou de reconnaissance de texte ne permettent pas l'insertion d'informations allographétiques. Prenons l'exemple d'un logiciel comme *Image Markup Tool* (Holmes et University of Victoria HMC 2010), qui permet d'associer une partie d'image à sa transcription et enregistre l'information dans des fichiers XML conformes à la TEI : il pourrait servir de base à des développements ultérieurs de sorte que les annotations de transcription puissent être structurées hiérarchiquement et correspondre au texte, c'est-à-dire qu'il faudrait pouvoir encoder des transcriptions comme des mots, voire des caractères (lettre ou signe abrégatif) et simplifier la transcription des abréviations. A l'heure actuelle, toutes les annotations sont de niveau égal, même si leur type diffère, et il n'y a pas d'imbrication des annotations et des zones¹⁶. De nouveaux projets en cours font naître de grands espoirs, tel Text-Image Linking Environment (TILE).

De l'autre côté, les logiciels d'OCR mériteraient d'être mieux intégrés, afin de limiter la perte d'information lors du passage des données graphiques primaires, analysées par le logiciel, à la sortie en ALTO ou TEI avec caractères Unicode, qu'il s'agisse de segmentation, ou, à l'intérieur d'un ensemble homogène, de reconnaissance de formes. Ces logiciels, même ceux qui reconnaissent certaines abréviations canoniques (*per*, *pre*,

¹⁶ On peut certes créer des catégories telles que « w » et « c » puis programmer une transformation pour inclure les éléments typés « c », mais c'est une étape supplémentaire, et la gestion des différentes zones devrait être adaptée en conséquence si l'on veut disposer d'un affichage par zones imbriquées.

que) n'offrent pas d'interface de modification manuelle pour enregistrer une analyse de type paléographique (résolution, typologie, allographes). La société Isako fournit un logiciel pour contrôler et modifier le résultat de la reconnaissance optique (logiciel OCRView), mais celui-ci, bien que couplé avec des logiciels livrant des fichiers ALTO, n'est pas adapté pour un ajout d'informations structurées, allographétiques ou non. Le projet européen Impact (Improving Access to Text)¹⁷ prévoit certes un nouvel outil de correction du texte reconnu, mais il n'est pas encore certain qu'il offrira la possibilité de récupérer la segmentation pour ajouter une information allographétique.

L'ensemble de ces outils ne peut donc pas, à l'heure actuelle, pallier directement la complexité d'un processus de transcription allographétique, ni prétraiter les images en repérant l'espace de la forme (segmentation qui évite aux chercheurs de doubler l'opération d'identification de la forme à celle, ô combien ingrate de sélection des formes). L'intégration de ces fonctionnalités des outils de production forme une étape indispensable afin de faciliter la transcription et l'établissement de liaison avec l'image de l'original.

5.2. Créer des outils d'analyse génériques

De l'autre côté, en aval de la production de fichiers à encodage allographétique, il faut améliorer les outils d'analyse, en offrant la possibilité de généraliser le questionnement et l'analyse. Cette dernière consiste à parcourir les bases de données paléographiques : en l'occurrence, en utilisant la TEI, cela revient à *parser* ou procéder à l'analyse syntaxique d'un fichier XML¹⁸.

Une interface générique doit être conçue pour paramétrer l'analyse des sources et obtenir, *in fine*, des informations chiffrées. Cette interface doit permettre non seulement de sélectionner les sources disponibles, mais aussi la nature de l'analyse : ainsi l'utilisateur doit être mesure de préciser sur quels allographes, quels paramètres, quelles abréviations porte son analyse.

Dans l'idéal, il faudrait être en mesure d'exploiter les fichiers existants selon leur pertinence et leur degré d'encodage. Pour cela, nous avons évoqué ci-dessus l'idée d'inscrire de façon formelle dans l'en-tête des fichiers le type de données exploitables. Cette formalisation n'existant pas à l'heure actuelle, une telle fonctionnalité ne peut être envisagée pour faire interopérer des fichiers hétérogènes. Si la situation s'améliore, le logiciel devrait également pouvoir tenir compte des ontologies utilisées et restreindre l'enquête aux sources pertinentes. Par exemple, une enquête à grande échelle sur

¹⁷ Le projet européen Impact (Improving Access to Text) rassemble 26 partenaires du privé et du public (bibliothèques nationales et de recherche) pour 4 ans (2008–2011), afin d'améliorer les techniques de reconnaissance de caractères et la mise à disposition du texte.

¹⁸ C'est avec une démarche de ce type que Florence Clavaud et moi avons travaillé : l'analyse se fait par XSLT, mais l'outil générique prévu au début n'a pas été développé jusqu'au bout.

l'influence de la fin de ligne ou des noms propres et des langues vernaculaires sur le travail des scribes (degré et type d'abréviation) ne peut se faire que sur des fichiers qui intègrent les notions d'abréviation, de lignes, de changements linguistiques et d'anthroponymes.

Nous ne considérons pas ici qu'il soit du ressort d'une telle application de préparer une édition électronique. Édition et encodage allographétique ont en effet des nécessités différentes, même si les deux peuvent se compléter réciproquement et ont en commun une longue histoire d'interdisciplinarité : l'attention accordée à l'une ne suffit pas nécessairement à l'autre¹⁹.

Il serait également bon de programmer des sorties sous un format directement importable ou exploitable par des logiciels standard d'analyse des données, c'est-à-dire non pas seulement sous un format HTML. Ainsi parviendrait-on à créer un environnement de travail où toutes les opérations automatisables seraient prises en charge par des outils adaptés, réservant au paléographe ce qui relève de son expertise : analyse des formes graphiques, élaboration des hypothèses, formalisation du questionnement et analyse des résultats à l'aide des outils de statistiques descriptives.

Que les principes présidant à l'établissement d'une édition ou description numérique dictent les possibilités d'exploitation du texte produit, cela est clair et a déjà donné lieu au jeu de mot « ce que tu (pré)vois est ce que tu obtiens » (Bradley 2005). C'est doublement vrai : aussi bien en amont de la préparation des sources qu'en aval, où l'on ne peut analyser qu'avec les outils à disposition, qu'il faut donc concevoir le plus largement possible. Une compréhension claire de ce que signifie aborder les écritures médiévales et transcrire est un préalable indispensable à une analyse paléographique assistée par ordinateur. Il nous semble que l'opération nommée « transcrire » signifie « décrire » dès lors que l'on s'intéresse aux formes graphiques et à l'image d'un texte autant qu'à son contenu. Aussi faut-il raisonner en termes de description et de structuration des informations, tant celles décrites que celles permettant de désigner les phénomènes décrits (ontologies, vocabulaires, descripteurs, granularité...). Une recherche en paléographie, menée avec des moyens réduits, nous a permis de faire la preuve de concept : une description allographétique minimale ouvre la voie à de nouvelles études sur l'histoire de l'écriture et des systèmes graphiques. Elle suffit à

¹⁹ Notre projet nous mène à une appréciation mitigée sur la complémentarité de l'encodage allographétique et de l'édition : il nous est apparu plus facile de réinsérer les abréviations et allographes étudiés *a posteriori*. Le problème a déjà été signalé par Robinson et Solopova : devoir veiller à trop de paramètres en même temps (abréviations, ponctuations, graphies) nuit à la qualité de la transcription. Dans notre cas, l'édition traditionnelle complète, avec appareil et notes, a précédé systématiquement l'encodage. La production des fichiers avec encodage allographétique n'exige pas l'analyse des noms de lieu et de personne ou de la syntaxe, analyse qui permet les capitalisations et ponctuation modernes que la machine peut restituer. L'étude des comportements scripturaux face aux noms et surnoms est un domaine pourtant fascinant qui peut inciter à étendre le champ d'encodage, surtout si les outils d'encodage rendent la tâche plus aisée.

éclairer des pratiques scripturales et des évolutions qui permettent de dater et d'identifier des écritures.

Pour avancer sur cette voie, il devient nécessaire de coopérer et de faire interopérer les bases de donnée paléographiques. Il faut donc, en amont, modéliser l'information pour pouvoir l'enregistrer dans sa complexité (normalisation et bonnes pratiques) et prévoir des outils qui favorisent une formalisation des choix d'encodage et une saisie aussi complète que possible, puis, en aval, des outils d'analyse. La coopération et le partage sont absolument nécessaires : d'une part entre les paléographes, qui pourront tirer grand profit de l'existence de données de comparaison, en évitant l'enfouissement et la déperdition de leurs propres données après aboutissement d'un projet ; d'autre part, entre les paléographes et les ingénieurs. Ceux-ci reconnaissent déjà l'intérêt de disposer de textes corrigés pour améliorer leurs algorithmes ; la mise en lumière de nouvelles règles correspondant aux différents types d'écritures pourra à son tour nourrir les paramètres exploités par les algorithmes, combinant segmentation, formes des lettres et règles de répartitions des allographes selon la position dans le mot.

Le travail conjoint des paléographes et des développeurs des outils d'analyse créera ainsi une masse d'informations paléographiques exploitable qui, en retour, pourra améliorer les algorithmes de reconnaissance des écritures en fournissant un corpus d'apprentissage. Seules la modélisation, la normalisation et l'élaboration de bonnes pratiques permettront de répondre aux défis industriels et scientifiques posés par les écritures manuscrites médiévales.

Bibliographie

- ABBYY : ABBYY FineReader 10. 2009. <<http://finereader.abbyy.com/>>.
- Anderson, Lisa, et al. *EpiDoc : Guidelines for Structured Markup of Epigraphic Texts in TEI*. Stoa Consortium, 2007. <<http://www.stoa.org/epidoc/gl/5/>>.
- André, Jacques. « Numérisation et codage des caractères de livres anciens. » *Document numérique* 7.3–4 (2007) : 127–142.
<http://www.cairn.info/article.php?ID_ARTICLE=DN_073_0127>.
- Aussems, Mark, et Axel Brink. « Digital palaeography. » *KPDZ* 1. 293–308.
- Beneventano, Domenico, et Sonia Bergamaschi. *Semantic search engines based on data integration systems*. International workshop on distributed agent-based retrieval tools. The future of search Engines' technologies. June 26, 2006 - PULA (CA – Italy). Cagliari : Center for Advanced Studies, Research and Development in Sardinia, 2006.
<<http://www.crs4.it/ict/dart06/slides/bergamaschi.pdf>>.
- Benskin, Michael. « The Hands of the Kildare Poems Manuscript. » *Irish University Review* 20 (1990) : 163–193.
- Bischoff, Bernhard. « La nomenclature des écritures livresques du IX^e au XIII^e siècle. » *Nomenclatures des écritures livresques du IX^e au XVI^e siècle : premier colloque international de paléographie latine, Paris, 28–30 avril 1953*. Eds. Bernhard Bischoff, Gerard Isaac Lieftinck

- et Giulio Battelli. Colloques internationaux du C.N.R.S. – Sciences humaines (Vol. 4). Paris : Édition du C.N.R.S., 1954. 7–14.
- Bozzolo, Carla, et al. « Les abréviations dans les livres liturgiques du ^{xv}^e siècle : pratique et théorie. » *La face cachée du livre médiéval : l'histoire du livre vue par Ezio Ornato, ses amis et ses collègues*. Actas del VIII coloquio del Comité internacional de paleografia latina (Madrid-Toledo, set.–oct. 1987) [Madrid, 1990, 17–27.] Rééd. Ezio Ornato. I libri di Viella 10. Roma : Viella, 1997. 555–565.
- Bradley, John. « What You (Fore)see is What You Get. Thinking about usage paradigms for computer assisted text analysis. » *TEXT Technology* 14.2 (2005) : 1-19.
http://texttechnology.mcmaster.ca/pdf/vol14_2/bradley14-2.pdf.
- Brink, Axel, Marius Bulacu, et Lambert Schomaker. « How much handwritten text is needed for text-independent writer verification and identification. » *Proceedings of 19th International Conference on Pattern Recognition (ICPR 2008)*, 8–11 December, Tampa, Florida. Los Alamitos : IEEC Computer Society, 2008.
<http://figment.cse.usf.edu/sfifilat/data/papers/WeBT6.2.pdf>.
- Bulacu, Marius, et Lambert Schomaker. « Automatic handwriting identification on medieval documents. » *14th International Conference on Image Analysis and Processing (ICIAP 2007). Proceedings. 11–13 September, Modena, Italy*. Los Alamitos : IEEE Computer Society, 2007. 279–284. <<http://www.ai.rug.nl/%7Ebulacu/iciap2007-bulacu-schomaker.pdf>>.
- Catach, Nina. « Graphémique. » *Lexikon der romanistischen Linguistik (LRL). Band I,1, Geschichte des Faches Romanistik, Methodologie : das Sprachsystem*. Eds. Michael Metzeltin, Christian Schmitt et Günter Holtus. Tübingen : Niemeyer, 2001a. 736–747.
- Catach, Nina. « Graphétique. » *Lexikon der romanistischen Linguistik (LRL). Band I,1, Geschichte des Faches Romanistik, Methodologie : das Sprachsystem*. Eds. Michael Metzeltin, Christian Schmitt et Günter Holtus. Tübingen : Niemeyer, 2001b. 725–735.
- Cerquiglini, Bernard « Une nouvelle philologie ? » *Philology in the Internet Era / Philologie à l'ère de l'Internet. International Colloquium / Colloque international*. Budapest : Eötvös Loránd University, 2000.
- Cottureau, Emilie. « La copie et les copistes français de manuscrits aux ^{xiv}^e et ^{xv}^e siècles. Etude sociologique et codicologique. » Thèse de doctorat. Université Paris 1 – Panthéon-Sorbonne, 2005.
- Criado Fernández, Luis, et Rafael Martínez-Tomás. « The problem of constructing general-purpose semantic search engines. » *Methods and models in artificial and natural computation. A homage to Professor Mira's scientific legacy*. Lecture Notes in Computer Science. Vol. 5601. Berlin, Heidelberg : Springer, 2009. 366–74.
- Derolez, Albert. *The Palaeography of Gothic Manuscript Books From the Twelfth to the Early Sixteenth Century*. Cambridge studies in palaeography and codicology. Vol. 9. Cambridge : Cambridge University Press, 2003.
- Ecole nationale des Chartes. *Conseils pour l'édition de textes médiévaux (Fascicule I, Conseils généraux. Fascicule II, Actes et documents d'archives. Fascicule III, Textes littéraires)*. Orientations et méthodes. 3 vols. Paris : Éd. du CTHS – École des chartes, 2001–2003.

- Flammarion, Hubert. « Une équipe de scribes au travail au XIII^e siècle : le grand cartulaire du chapitre cathédral de Langres. » *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde* 28 (1982) : 271–305.
- Flammarion, Hubert. *Cartulaire du chapitre cathédral de Langres*. Nancy : ARTEM, 1995, 449 p. ARTEM Atelier de recherches sur les textes médiévaux. 2^e ed. Turnhout : Brepols, 2004. 525 p.
- Frioli, Donatella. « La ‘Grammatica della Leggibilità’ nel manoscritto cisterciense. L’esempio di Aldersbach. » *Liturgie und Buchkunst der Zisterzienser im 12. Jahrhundert : Katalogisierung von Handschriften der Zisterzienserbibliotheken*. Ed. Charlotte Ziegler. Frankfurt am Main : Peter Lang, 2000. 17–47.
- Frioli, Donatella. « Per una storia dello scriptorium di Reichersberg. Il prevosto Gerhoch e i suoi ‘segretari’. » *Scrittura e civiltà* 23 (1999) : 177–212.
- Garand, Monique-Cécile. « Copistes de Cluny au temps de saint Maieul (948–994). » *Bibliothèque de l’École des Chartes* 136.1 (1978) : 5–36.
- Gasparri, Françoise. *L’écriture des actes de Louis VI, Louis VII et Philippe Auguste*. Hautes études médiévales et modernes. Vol. 20. Paris, Genève : Minard, Droz, 1973.
- Gatos, Basilis, Nikolaos Stamatopoulos, et Georgios Louloudis. *ICDAR2009 Handwriting Segmentation Contest*. 10th International Conference on Document Analysis and Recognition. Athens : Institute of Informatics and Telecommunications. National Center for Scientific Research « Demokritos », 2010.
<<http://users.iit.demokritos.gr/%7Ebgat/HandSegmCont2009/HandSegmCont2009.pdf>>.
- Gentleman, Robert, Ross Ihaka, et R Development Core Team. R. version 2.5.0 ed : The R Foundation for Statistical Computing, 2007.
- Guillot, Céline, et al. « Constitution et exploitation des corpus d’ancien et de moyen français. » *Corpus* 7 (2008). <<http://corpus.revues.org/index1495.html>>.
- Haugen, Odd Einar (ed.). *The Menota handbook : Guidelines for the electronic encoding of Medieval Nordic primary sources*. version 2.0 ed. Bergen : Medieval Nordic Text Archive, 2008. <<http://www.menota.org/guidelines>>.
- Heiden, Serge, Céline Guillot, et Alexei Lavrentiev. *Manuel d’encodage XML-TEI des textes de la Base de Français Médiéval*. 2002–2008.
<http://ccfm.ens-lsh.fr/IMG/pdf/Manuel_Encodage_TEI.pdf>.
- Heinemeyer, Walter. *Studien zur Geschichte der gotischen Urkundenschrift*. Archiv für Diplomatik. Beiheft 4. Köln : Böhlau Verlag, 1982.
- Hofmeister, Wernfried, Andrea Hofmeister-Winter, et Georg Thallinger. « Forschung am Rande des paläographischen Zweifels : Die EDV-basierte Erfassung individueller Schriftzüge im Projekt DAMaIS. » *KPDZ* 1. 261–292.
- Holmes, Martin, et University of Victoria HCMC. *Image Markup Tool. Tool for annotating images using TEI*. version 1.8.1.7 ed. 2010. <http://tapor.uvic.ca/mholmes/image_markup/>.
- IMPACT. Den Haag : Koninklijke Bibliotheek, 2008–2010.
<<http://www.impact-project.eu/index.php>>.
- KPDZ 1 : *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Eds. Malte Rehbein, Patrick Sahle et Torsten Schaßan. Schriften des

- Instituts für Dokumentologie und Editorik 2. Norderstedt : Books on Demand, 2009. En ligne : <[urn:nbn:de:hbz:38-29393](http://nbn-resolving.org/urn:nbn:de:hbz:38-29393)>, <<http://kups.ub.uni-koeln.de/volltexte/2009/2939/>>.
- Lavrentiev, Alexei. *Proposal for XML markup of Old French text corpora*. The Princeton Charrette Project. Princeton, 2002.
<<http://www.princeton.edu/%7Elancelot/ss/media/docs/Transcription-Proposal.doc>>.
- Lavrentiev, Alexei. *Systèmes graphiques de manuscrits médiévaux et incunables français : ponctuation, segmentation, graphies : actes de la journée d'étude de l'ENS LSH, 6 juin 2005*. Langages. Vol. 3. Chambéry : Université de Savoie, 2007.
- Leydier, Yann, Jean Duong, et Asma Ouji. *Ulysse 0.3g09*. 2006–2009.
<<http://liris.cnrs.fr/graphem/?p=73>>.
- Mane, Laure. *TELplus. WP3 Task 1 – Indexing for usability. A prototype of semantic full-text search engine indexing multilingual OCRed corpus from European digital libraries. Feasibility assessment report*. Den Haag : The European Library, 2010. <http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/documents/TELplus_D3.3_04012010.pdf>.
- Mazziota, Nicolas. « Traiter les abréviations du français médiéval. Théorie de l'écriture et pratiques d'encodage. » *Corpus 7* (2008). <<http://corpus.revues.org/index1517.html>>.
- McGillivray, Murray. « The Cotton Nero A.x. Project. » 1994–2010.
<<http://people.ucalgary.ca/Scriptor/cotton/>>.
- McGillivray, Murray. « Statistical analysis of digital paleographic data : what can it tell us ? » *TEXT Technology* 14.1 (2005) : 47–60. <http://texttechnology.mcmaster.ca/pdf/vol14_1_05.pdf>.
- Meyer, Wilhelm. *Die Buchstaben-Verbindungen der sogenannten gothischen Schrift*. Abhandlungen der königlichen Gesellschaft der Wissenschaften zu Göttingen, Phil.-hist. Klasse. Vol. Neue Folge, 1,6. Berlin : Weidmannsche Buchhandlung, 1897.
- Millares Carlo, Agustín, et José Manuel Ruiz Asencio. *Tratado de paleografía española*. 3^e ed. 3 vols. Madrid : Espasa-Calpe, 1983.
- MUFI : *Medieval Unicode Font Initiative*. *MUFI character recommendations*. Bergen, 2009.
<<http://www.mufi.info/specs/>>.
- Muzerelle, Denis. « Graphem for Dummies. » *The Manuscript Triangle France-England-Scandinavia. 1100-1300*. Bergen : University of Bergen, 2009.
<<http://www.uib.no/filearchive/graphemfords.pdf>>.
- Nicolaj, Giovanna. « Alle origini della minuscola notarile italiana e dei suoi caratteri storici. » *Scrittura e civiltà* 10 (1987) : 49–82.
- Oeser, Wolfgang. « Beobachtungen zur Strukturierung und Variantenbildung der Textura. Ein Beitrag zur Paläographie des Hoch- und Spätmittelalters. » *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde* 40 (1994) : 359–439.
- Oeser, Wolfgang. « Beobachtungen zur Differenzierung in der gotischen Buchschrift. Das Phänomen des Semiquadratus. » *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde* 47–48 (2001–2002) : 223–83.
- Pantarotto, Martina. « La scrittura delle carte bresciane nel sec. XII. » *Scrineum – Rivista* 3 (2005) : 1–20. <<http://scrineum.unipv.it/rivista/3-2005/pantarotto.pdf>>.
- Petrucchi, Armando. « Censimento dei codici dei secoli XI–XII. » *Studi medievali*, Ser. 3, 9.2 (1968) : 1115–1194. <<http://dida.let.unicas.it/links/didattica/palma/testi/petrucchi1.htm>>.

- Richard, Jean. *Les ducs de Bourgogne et la formation du duché du XI^e au XIV^e siècle*. Publications de l'université de Dijon 12. Paris : Les Belles Lettres, 1954.
- Richard, Jean. « La chancellerie des ducs de Bourgogne de la fin du XII^e au début du XV^e siècle. » *Landesherrliche Kanzleien im Spätmittelalter. Referate zum VI. Internationalen Kongreß für Diplomatie, München 1983*. Vol. 1. Ed. Gabriel Silagi. München : Ardeo, 1984. 381–413.
- Robinson, Peter, et Elizabeth Solopova. « Guidelines for the transcription of the manuscripts of the *Wife of Bath's* Prologue. » *The Canterbury Tales Project Occasional Papers*. Vol. 5. Ed. Norman Blake. Oxford : Office for Humanities Communication, 1993. 19–52.
<<http://www.canterburytalesproject.org/pubs/transguide-MI.pdf>>.
- Rumble, Alexander. *The palaeographical material in the C11 Database [Dating and describing eleventh-century vernacular script]*. MANCASS C11 Database Project. Manchester : Manchester Centre for Anglo-Saxon Studies, [2004].
<<http://www.arts.manchester.ac.uk/mancass/C11database/data/PalaeogIntro.pdf>>.
- Salaün, Jean-Michel. « Web, texte, conversation et redocumentarisation. » *JADT 2008 : actes des 9es Journées internationales d'Analyse statistique des Données Textuelles, Lyon, 12–14 mars 2008 : proceedings of 9th International Conference on Textual Data statistical Analysis, Lyon, March 12–14, 2008*. Eds. Serge Heiden et Bénédicte Pincemin. Lyon : Presses universitaires de Lyon, 2008. <<http://jadt2008.ens-lyon.fr/spip.php?article197>>.
- Stokes, Peter A. « Computer-aided palaeography, present and future. » *KPDZ* 1. 309–338.
- Stutzmann, Dominique. « Écrire à Fontenay. Esprit cistercien et pratiques de l'écrit en Bourgogne (XII^e–XIII^e siècles). » Thèse de doctorat. Université Paris 1 Panthéon-Sorbonne, 2009a.
- Stutzmann, Dominique. « La sobriété ostentatoire : l'esthétique cistercienne d'après les manuscrits de Fontenay. » *Culture et patrimoine cisterciens. Colloque du vendredi 12 juin 2009*. Vol. 4. Parole et Silence / Cours, colloques et conférences. Paris : Collège des Bernardins, 2009b. 46–86.
- Tassin, René Prosper, et Charles-François Toustain. *Nouveau traité de diplomatique, où l'on examine les fondemens de cet art [...] par deux religieux bénédictins de la Congrégation de S. Maur*. 6 vols. Paris : G. Desprez, 1750–1765.
- TEI Consortium. « TEI P5. Guidelines for electronic text encoding and interchange ». TEI consortium, 2007–2010. <<http://www.tei-c.org/release/doc/tei-p5-doc/html/>>.
- TILE : *Text-Image Linking Environment*. Maryland Institute for Technology in the Humanities, 2009. <<http://mith.info/tile/>>.
- Tomasi, Gilbert, et Roland Tomasi. « Approche informatique du document manuscrit. » *KPDZ* 1. 197–218.
- Torrens, Maria Jesus. « La paleografía como instrumento de datación. La escritura denominada “littera textualis”. » *Cahiers de linguistique hispanique médiévale* 20 (1995) : 345–380.
- Tzadok, Asaf, et Eugeniusz Walach. « Adaptive OCR for Books. » Armonk (NY) : International Business Machines Corporation, 2009.
<<http://www.freepatentsonline.com/7627177.html>>.
- Uitti, Karl D. *Charrette Project SGML Codes*. 1997.
<<http://www.princeton.edu/lancelot/ss/materials.shtml#ms-transcriptions>>.

- Zaluska, Yolanta. *L'enluminure et le scriptorium de Cîteaux au XIIe siècle*. Studia et documenta. Vol. 4. Abbaye de Cîteaux (Saint-Nicolas-les-Cîteaux) : Cîteaux Commentarii cistercienses, 1989.
- Zamponi, Stefano. « La scrittura del libro nel Duecento. » *Civiltà comunale. Libro, scrittura, documento. Atti del Convegno (Genova, 8–11 novembre 1988)*. Nuova Serie, Vol. 29, fasc. 2. Atti della Società Ligure di Storia Patria. Genova : Società Ligure di storia patria, 1989. 317–346.

Agendas for Digital Palaeography in an Archaeological Context: Egypt 1800 BC

Stephen Quirke

Abstract

Handwriting raises issues alive in archaeological debates, philosophical and historical. In turn, by their extreme fragmentariness, the earliest archaeological manuscripts could generate usefully different questions for the field of palaeography. Here, digitisation offers new common ground for the separate disciplines in the study of the past. For current archaeological discussions of structure and agency, manuscripts pose the act of writing, between social and individual. For debates over literacy and power in part-literate societies, an archaeological hoard of manuscript fragments offers opportunities to assess our chances of knowing, for one time and place, how many writings and writers. The largest earliest group of writing on papyrus-paper comprises several thousand small fragments from Lahun in Egypt (about 1850–1750 BC). Traditional methods of recording similarity and difference across the collection can now be accelerated to a point of qualitative change, by applying image-matching software. This paper considers the potential of computer-aided palaeography for generating new research agendas.

Zusammenfassung

Schrift als Kulturphänomen ist ein zentraler Gegenstand archäologischer Debatten, in philosophischer wie in historiographischer Hinsicht. Dabei können die frühesten archäologischen Handschriftenfunde in ihrer extremen Fragmentierung Anstoß für weitreichende Forschungsfragen innerhalb einer allgemeinen Paläographie geben. Digitalisierung bietet hier den unterschiedlichen Disziplinen eine gemeinsame Grundlage für die Erforschung der Vergangenheit. Innerhalb der gegenwärtigen archäologischen Debatten um Struktur und Akteur (*agency*) positionieren Handschriften den Schreibakt zwischen sozialem und individuellem Handeln. Für die Diskussion über den Zusammenhang von Schriftlichkeit und Macht in teilschriftlichen Gesellschaften können lokal und zeitlich dichte archäologische Funde von Handschriftenfragmenten erstmals präzisen Aufschluss über die Anzahl von Schreibern und Schriftstücken an einem bestimmten Ort zu einer bestimmten Zeit geben. Bei den Fragmenten aus dem ägyptischen Lahun (ca. 1850–1750 v.Chr.) handelt es sich um die größte Ansammlung frühester Papyrusschriftstücke. Die traditionellen Methoden zur Feststellung von Ähnlichkeiten und Unterschieden innerhalb einer kompletten Sammlung können nun mit Hilfe

von digitalen Bildvergleichsverfahren auf einem qualitativ neuen Niveau angewandt werden. Dieser Beitrag möchte Potential und Perspektiven einer computergestützten Paläographie für die archäologische Forschung aufzeigen.

1. Handwriting in Debate

1.1. Structure and Agency

In the twentieth century division of labour for studying pasts, philological historians parted company from archaeological fieldworkers (Andrén). Archaeological debates on structure and agency have missed ancient handwriting as a case-study of socially constrained individual actions. Here, ‘action’ may be conceived as the interface where two dimensions come into existence: mutually constituting agents, and structures they (the agents) live, in a dialectic ruled by conflicts expressing contradictory interests between and within each stratum.¹ In posture and body movement, writers at work experience as individuals instances of a shared activity: each may develop and express their identity in writing style, in relation to acquired skills, physical capacities, and degrees of self-control. In different historical contexts, writers may produce their materials themselves, or receive from suppliers with separate production spaces each or all of the range: ink, writing support, writing-tools and containers.

Just as speakers of a natural language do not passively reflect or label a world of things, but rather move within and contribute to that language as a conception of life (Ives), so too writers succeed or fail in communicating not in isolation, but within forms accepted in the community of writers and readers at their time, and within its traditions of learning to write and to read. Therefore handwriting is natural-historical, imbued with individual and social dimensions, and accordingly changes over time. This temporal flow creates for palaeography two of its privileged tasks: to assess the dates of manuscripts, and to re-read forgotten writing (Hofmeister et al. 268). For dating, palaeographers may find changes over time gradual or staccato; to echo the terms of Mark Stansbury, gradual change perhaps fosters evolutionary interpretation, whereas more abrupt change may encourage taxonomic periodisation. Ancient Egyptian handwriting follows both paths over the long duration. In the second millennium BC, Egyptian cursive handwriting evolved gradually for two or more centuries, before then being revised at a fixed point, as if by decree from a centralised point of learning (Roccati). Whatever the institutional channels, handwriting revision may be observed in the mid-nineteenth, fifteenth, and thirteenth centuries BC (Möller).

For the specific space and time of script learning in Egypt, secure archaeological evidence remains elusive. Like the library of Alexandria (Butler), the ancient Egyptian

¹ Callinicos 184–188, from commentary on Giddens by Margaret Archer, without the part about God.

‘school’ remains a mirage conjured by one line at the start of a single literary composition of the nineteenth century BC, the Teaching of Khety:²

Beginning of the teaching made by the man of Tjaru (?)
 the hymn-singer (?) called Khety for his son called Pepy
 In the very time of sailing south to the Residence
 to place him in the teaching-chamber of writings
 among the children of officials, of the foremost of the Residence

By re-translating “teaching-chamber of writings” as “writing school”, Egyptologists have equated ancient teaching and modern education, in a manner contested even for nineteenth century Egypt teaching (Mitchell). In the question of script learning, against the images of Greek, Roman and European Renaissance mystifications, it may be noted that the ancient Egyptian scripts provide a perfect solution for communicating the structure of the ancient Egyptian language. Their combination of specific or generic image-signs with one-, two- and three-consonantal sound-signs removes the risk of confusing the many similar-sounding words generated from its core trilateral roots. In Egypt, development of an alphabet brings no discernible progress for literacy, and is a regressive step for ability to convey the language. In the early first millennium AD, some five centuries after its first appearance in the country, the Greek alphabet was adapted to write the Egyptian language as the Coptic script (Bosson and Aufrère). Though doubtless a more international medium, the alphabet seems a major setback in conveying Egyptian language, as may be experienced still today by anyone who learns Egyptian in hieroglyphic and Coptic scripts at the same time.

With no production-spaces, and few ancient writing-kits preserved, the manuscripts and their signs in pigment and an unidentified binding medium remain the direct evidence for transmission of writing and reading over generations.

1.2. Literacy and Power—Assumptions—Datasets

As formalised communication deployed in control of material and intangible resources, writing is as much an exercise of power as any other act of speech or tooling. Foucault investigated institutions of power across temporal blocks of *episteme* “knowledge formation”, a term more digestible to most of academic society than its Marxist original, in socio-economic history, with the vocabulary of modes of production and social formations (cf Jameson). In general, such longer-term structural history, like philosophy, risks shipwreck on the exacting details of short-term micro-history, of the kind that

² Translation following the synoptic edition Helck; the word here translated “hymn-singer” has also been interpreted as a rare personal name Duau, but is attested in a list of officials predominantly comprising temple staff, UC32194, Collier and Quirke 2006, 100–101.

gave philology its reputation for pedantry (cf Gran). Yet, in their practices for managing myriad fragments, philologists have discretely sustained their own philosophy, in part by declining to theorise either their own actions, or the fragmentariness of the record.³ Disciplines addressing remoter pasts have imposed rough assumptions concerning low literacy rates, relations between power and literacy, and the presence, age and gender of the literate in ancient landscapes.⁴ The internally contradictory detail and quantity of nineteenth century documents help check our assumptions, not as direct illustrations, but as reminders to return more open-mindedly to the ancient and differently fragmentary sources.

The AD 1897 census for Egypt, in the second decade of British military occupation, gives minimal levels of literacy for women in all regions, and for men outside the urban governorates (*Recensement général*):

	Total	Governorates	Lower Egypt	Upper Egypt
men	8.8%	22.6%	8.4%	5.9%
women	0.6%	0.57%	0.01%	0.01%

Table 1. Literacy in Egypt AD 1897.

Even leaving aside questions of criteria and methodology in such a census, such numerals need more than block-reading. Four letters from an archaeological archive of the time indicate how such literacy levels might operate in practice at rural margins.

1. On March 8th 1884, in the Nile Delta, the archaeologist Flinders Petrie recorded how he received an Arabic letter from Ali Jabri, his illiterate organiser at the Bedouin village beside the Giza pyramids (Quirke 2007: 96): “Next morning one of the boys here who can read and write (what a treasure a scribe on the premises is I cannot tell) told out in a long singsong drawl the contents of Ali’s letter.”
2. In November 1891, Petrie recruited five workforce supervisors at al-Lahun village, on his way south to excavate at al-Amarna. One was literate, one learning (Drower 81): “A strapping lad of about 20 is Abdallah, who has the advantage of reading & writing (as Misid also a little).”
3. On January 9th 1905, Petrie wrote to his wife Hilda (Quirke 2007: 81): “Our men have not had a single letter from Quft. Aly abd er Rahim desires you to tell Ib. that they are all well here, and wishes him to write this to Quft. They have sent several letters to Quft themselves.”

³ The relation between implicit and explicit might be added as a dimension to the diagram of knowledge construction in Cartelli and Palma 132: Fig. 2.

⁴ Shubert, contesting these assumptions on gender.

4. On January 17th 1906, excavating at Tell el Retaba, Hilda noted of friendly hosts at a Nile Delta station (Quirke 2007: 88): “Two of them came over for a lesson in archaeology next day, and they write us English letters in answer to my Arabic ones, and are very obliging in procuring bread for us.”

In these documents, individual and collective practices cannot easily be reduced to a binary literate/non-literate divide. Reading and writing interweave, deploying devices of literacy, orality, and, wherever individuals or groups listen, aurality (Coleman 1–33). Moreover, here English views of Egypt, ancient and modern, can be reversed by reading from Egyptian perspectives. For, on site, a nineteenth-century census-taker might have marked the London-based dig-director as illiterate, despite his economic power. Petrie knew enough Arabic to converse with excavators, and write names of people and places, but this might not qualify him as literate to the standards of a local government officer.

In Egypt 3000–1000 BC, eternity and modernity each have their medium and script (Assmann). For monuments that project life into eternity, sacralising space, the medium is stone or metal, and the script comprises signs with the same proportions as formal art (Fischer). In contrast, the writing of ‘modernity’, defined as the contemporary horizon of each living generation, concerns letters and accounts, and then, increasingly over time, literary, technical and religious compositions. The dominant medium on this horizon is a paper made from strips of papyrus-reed; its script comprises the signs from the script of eternity, written with a *Juncus maritimus* reed, which created more fluent and, soon, cursive forms. The writer touched the reed tip into a water-pot and onto a cake of pigment with the unidentified binding medium presumed often to be gum arabic (Tait and Leach). The dominant pigment was carbon black, the optional highlighting pigment red ochre. Writers may have obtained supplies of ready-made paper and pigment, but the sources, regularity and chains of supply are undocumented. Study of longer manuscripts confirms the written evidence for production of rolls formed of twenty joined sheets. Sheet-joins are either at regular intervals and very fine, sometimes invisible, or at irregular points and coarse, so easily detectable. Fine joins are presumably the work of professional book-producers, while rougher joins indicate points where the writer has added a sheet for extra space. Book-supply and paper-use in this world before the codex are therefore rather different to those in modern paper-production, of pages laid into quires and then bound to books. From recurrent differences in sheet-joins, it seems that Bronze Age Egyptian writers received, not page-like sheets, but the ready-made book-rolls; from a roll any fraction may be cut or torn, for separate use e.g. as letters, or to be added to book-rolls to form a longer roll. Subtracting and adding create a different inscriptional field of practice, to that of the normative printed books in modern times.

Dwarfed by the more abundant preservation of sacred inscription on stone, three larger groups of manuscripts stand out as datasets for research:

Early Bronze Age 2650–2000 BC	Abusir temple business papers (Posener-Kriéger)
Middle Bronze Age 1850–1750 BC	Lahun temple business papers (Kaplony-Heckel 1971), town miscellaneous (Collier and Quirke 2002, 2004, 2006)
Late Bronze Age 1300–1050 BC	Deir el-Medina business papers, miscellaneous (Valbelle)

The Lahun town papyri represent the earliest miscellaneous group, offering a random sample ideal for analysis. Preserved in the Petrie Museum at University College London (UCL), they form the basis for this review.

2. Lahun Papyri as Dataset

2.1. Place and Time

In the early nineteenth century BC, whoever planned kingship temples decided to locate the pyramid complex for king Senusret II at the entrance to Fayoum, near modern al-Lahun (Grajetzki). Alongside the Valley Temple of this complex, a kilometre east of the pyramid itself, at or about the same time, a new town was laid out on an orthogonal plan (Petrie 1891: pl. 14). Its name seems to have been Hetep-Senusret “Peace of Senusret” (Horváth 2009a), and its main block measures about 250 by 250 metres (500 by 500 ancient Egyptian cubits), with ten palatial houses on the north, upwind of medium and small-sized houses. An additional strip of housing and, perhaps, administrative buildings forms a contiguous western sector closer to the Valley Temple.

In 1889 Egyptian teams recruited from Madinat al-Fayum and al-Lahun cleared the town-site, in two seasons directed by Flinders Petrie, whose primary aim was to retrieve the town-plan and representative finds of its period (Petrie 1890, 1891). The archaeological generation of Petrie did not yet use survey grid or stratigraphic sections to record the horizontal and vertical relation of finds; instead, in his spring 1889 season he assigned letters to ‘Ranks’, meaning blocks of houses between streets. Unsurprisingly, then, for that date in the history of archaeology, find-places were only recorded for three of the five larger groups (those containing more than five separable items): Lots I+II from Rank C head, Lot III from Rank B head, Lot IV probably from middle block of Rank N; the find-spots for Lots VI and LV are not recorded (Collier and Quirke 2006). Surprisingly perhaps for current disciplinary archaeology, the prompt to record specific find-places came from a philologist, Francis Llewellyn Griffith (Collier). It is regular practice to lament the loss of exact provenance within the site, although the absence of information possibly reflects an undramatic scatter across the entire town—in itself

a caution on Egyptological low literacy estimates at least for these more urbanised landscapes. A new survey is now being directed by Zoltan Horváth of the Museum of Fine Arts Budapest (Horváth 2009b). However, the 1990s excavation team of the late Nick Millet found the site to have been almost entirely stripped of its bricks since the 1889 clearance (Frey and Knudstad).

Entrusted with publication of the papyri, Griffith thanked Petrie in print for keeping separate the material in batches as it was presented to him (Petrie 1891: 47). At least eighty 'Lots' were handed over to Petrie from across this 250x250m town, amounting to a productively random sample from a century of written output. The 1898 edition by Griffith of 65 "of the best" made the collection of Lahun papyri preserved at UCL famous, as they include the earliest legal, literary, mathematical, medical papyri, and the only veterinary manuscript from Bronze Age Egypt (Collier and Quirke 2004). Since the 1898 edition, museum staff have rescued the papyri from two World Wars, and a fire started by conservation materials (Kaplony-Heckel 1980: 293 n. 2), causing soot-damage to some frames but, fortunately, no material loss. Whether the collection can continue to survive their present Antimuseum home in a converted stables-building remains an open question for the university. By 1990, after all the moves over twelve decades of study and storage, half of the fragments had no Petrie batch-number, complicating efforts to calculate how many manuscripts, and how many writers, are present in this exceptional haul of writing. In sum, they present a prime papyrological jigsaw puzzle for archaeologists and historians of the Bronze Age.

2.2. Change in the nineteenth Century BC

In order to appreciate the particular challenge of Lahun handwriting, a later change may be compared for contrast. In the Hellenistic Period, the reed and language of Kemet were displaced by Greek script and diagonally-cut *phragmites* rush (Tait and Leach). The *phragmites* rush pen may be labelled Greek in Egyptology and papyrology, but could derive from earlier first millennium BC practice in Nubia or Assyrian west Asia. Whatever the origin of the tool, Greek and Egyptian scripts became associated with different writing-kits and materials; pigment analyses on third century BC bilingual documents have demonstrated use of lead inks for the Greek script, beside carbon-black pigment brush-stroke for Egyptian demotic (Delange, Grange, Kusko, Menei). By the first century AD, pigment and writing-tool changed from Egyptian carbon and reed to Greek lead and rush even for religious compositions in hieratic, the older cursive (Quaeghebeur). The new writing-tool favours angular in place of rounded signs; sometimes it is difficult to determine which writing-tool was used, because even reed-users reproduce the angular signs that derive from writing with the cut rush. Change in writing style can be correlated here with change in instrument.



Figure 1. Lahun papyrus fragments UC32137I with angular handwriting (before 1850 BC), and UC32171J, with rounded handwriting (after 1850 BC).

No new writing equipment is available to explain the change in the reverse direction, from angular to rounded, in the aesthetic of writing practice during the nineteenth century BC. In the Lahun temple accounts, angular signs give way to rounded remarkably abruptly, in the decade preceding the reign of king Amenemhat III, about 1850 BC (Luft 21).⁵ At the same time, the material as well as the verbal culture of Egypt underwent major changes in every area, leading Egyptologists now to distinguish sharply between a late and an early Middle Kingdom. The inclusion of handwriting in this cultural revolution calls for particular investigation.

3. Handwriting and Orthography as Egyptological starting Frames—Potential, Limitation

As in the far larger field of medieval studies (Stokes 316), both handwriting and orthography provide Egyptologists with ground for dating manuscripts. Orthography has been explored more rarely, though with particular success in dating the earliest

⁵ Compare UC32137I angular with UC32171E and J, rounded, on frame of fragments Fig.1.

literary manuscripts (Dévaud). More often, researchers have targeted handwriting. A century ago, Georg Möller compiled a palaeographical reference book for Egyptology, isolating single signs by tracing from photographs or printed facsimiles of manuscripts, and including a column for the Lahun papyri (Möller). From these three volumes on hieratic (excluding the more cursive demotic), Alan Gardiner derived his own standard sign-list of hieroglyphs in his 1927 *Egyptian Grammar*—here by inversion the cursive everyday script became the anchor of the sacred script. Since Möller, improvements on individual readings have tended to come from new evidence in clear hieroglyphic forms, for example in the name of the town quarter nearer the pyramid-complex, Sekhem-Senusret rather than the Griffith reading Ankh-Senusret (Gunn). The Möller palaeography is intended to date undated papyri, prior to identifying individual hands. Continuing research in this tradition includes study of writings of *pa* “the” in twelfth century BC correspondence (Janssen).

Following conservation of the Lahun papyri by Bridget Leach and, after rediscovery of the smallest fragments in 1994, Renee Waltham, in 2002–2006 Mark Collier and myself published the entire collection in three volumes of transliteration and translation with CD-ROM of colour scans. From the start of our work in 1991, we found that ancient format had dictated form of sign to such an extent that nearly every inscribed fragment larger than two centimetres could be assigned to a particular type of content.⁶ In effect, we felt, our content categories emerged from the signs themselves as deployed across a page or a book-roll of different dimensions, further defined by cultural choices in the presentation of certain contents. In order of quantity, the final categories were accounts, letters, and the more miscellaneous array of legal, literary, religious, medical, and mathematical. Within book-rolls, accounts rolls tend to be taller (full height roll, around 30 cm) than literary (half-heights or quarter-heights, 15–16 or 7–8 cm); heights vary for technical treatises such as the ‘gynaecological’ papyrus and other healing books. Legal documents tend to a layout between accounts and reports, with sign forms often close to those found in the literary and accounts rolls, whereas letters are demarcated by vertical columns introducing or sometimes framing a core of horizontal lines. Titles are not used for any content category. However, shorter ‘page’-documents may have an address (letter) or contents (legal) line in the patch on the back that would face outwards after rolling and tying, sometimes secured by sealing-string and mud with stamp-seal impression, sometimes by improvisation as with a fish-hook in one instance. All layout varies within a culturally-determined physical field, the space of the lap of writer seated on the ground with legs folded.

Study since our edition has produced new joins, sometimes confirming the categories, sometimes putting them in doubt. Within the accounts fragments, it has been possible to redefine a group with distinctive large, rounded signs (Quirke 2007). Across and

⁶ For layout script-styles, not secure signs of writer individuality, see Hofmeister et al. 277.

against our categories, one “religious” fragment can now be identified as part of the same manuscript as a second fragment categorised “literary religious” (Müller). This identification confirms the impact of print publication, with CD-ROM for colour information; a global army of readers can tackle problems of reading and identification of single manuscripts and individual hands, preparing discussion on broader issues such as literacy. At this point we may ask, considering that wider research access, which or indeed whether new tools are needed, and what ramifications would they bring. Palaeography has been said to tackle three main questions: when, where, and “were these different things written by the same person” (Stokes 310, cf Aussems and Brink 293). If we take as our level of resolution, not political history, but period of material culture, then the AD 1889 excavation can answer when and where as “Lahun, 1850–1750 BC”, leaving the third question, how many hands, how many manuscripts. Arianna Ciula (219) emphasises how “humanities computing methods can assist in making explicit” processes of the palaeographical. Precisely to explain my Lahun project objectives to colleagues in computing/engineering, I articulate the third question as three objects for identification:

1. similar handwritings across all fragments, to map (a) dispersed manuscripts, (b) writers or writer-groups present in more than one ‘Lot’;
2. similar handwritings within each ‘Lot’ with particular attention to:
 - intra-category: similar handwritings within a single category as defined by the combinations of formal features (e.g. ‘letters’);
 - inter-category: similar handwritings across different content categories, particularly those with different layout on the page;
3. range of difference within a single manuscript to establish the parameters of difference against which to measure the apparent similars of tasks 1–2.

Underlying the three tasks are the “one writer, many hands”, and “one hand, many writers”. Even in the larger field of medieval manuscript studies, the “stilistische Schwankungsbreite von Schreiberhänden” has been considered a crucial factor not yet adequately researched, when particularly more skilled writers might deploy more than one handwriting style and even ductus.⁷ Conversely, as Stokes notes (317), “the very question of scribal identity depends [...] on the assumption that the handwriting of no two persons is the same, and yet this assumption is not normally questioned by palaeographers”. He finds reassurance in forensic science, where handwriting proves consistent with writer individuality, but also warns that this varies by degree of training of the writer or group of writers. At the greater distance of Bronze Age Egypt, research can afford assumptions even less, and needs new tools and their horizons.

⁷ Hofmeister et al. 276–277 with n. 26; Stokes 315 on the typology of copying, with both imitation and influence of exemplars.

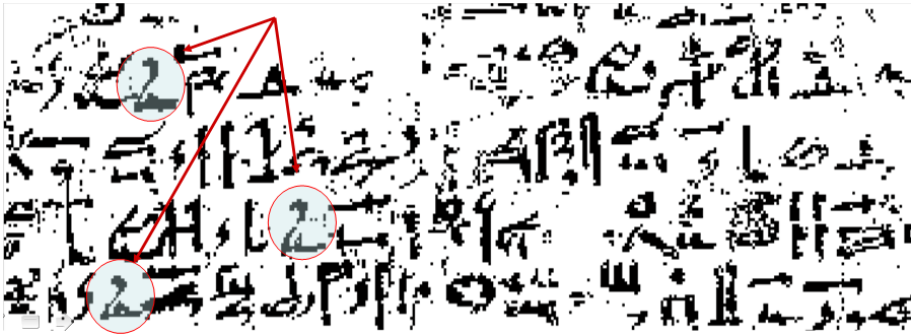


Figure 2. Tania Stathaki trial for identification of *alif* single-consonant sign in Lahun papyrus.

4. Computer-aided Palaeography—Trials, Potential, Agendas

From 2004, initial Imperial College London MSc projects supervised by Anthony Constantinides, demonstrated how far computer-aided palaeography could accelerate traditional tasks for the Lahun papyri. However, limitations of smaller-scale projects also became clear by the third year, anticipating the observation by Arianna Ciula on the “difficulty of maintaining a project, which was never formally funded” (Ciula 232). Despite successful character recognition, serious obstacles remained, including the need for higher resolution scanning (cf Fig. 2), and the quantity of tiny fragments. A third, more intractable obstacle, was interdisciplinary time threshold: two disciplines need time to develop deeper engagement than dual-supervised dissertations provide. Resolution is being addressed in-house by rescanning, under the supervision of Tim Weyrich at UCL Computer Science, with his new software to overcome distortions in flat-bed scanning, in particular for glass-mounted fragments. The other obstacles require longer-term approaches, for which Anthony Constantinides, his colleague Tania Stathaki, Tim Weyrich and myself envisage a postdoctoral research project, combining the philological and engineering resources available in London. Much as I needed to re-articulate the task of identification, I needed a translation back to me, as non-digital philologist. Stathaki articulated the computing tasks in these terms:

The primary aim of identification is to locate automatically and identify signs within sequences of signs.

Shape representation: Create a set of features that adequately encode the characteristics of the symbol.

Shape matching: Endow the feature space with an appropriate similarity/dissimilarity metric.

Re-scanned, the two hundred frames with several thousand fragments can then provide an archaeological dataset to rise to the challenge: “Are there new possibilities for manuscript research to be discovered that had not been possible before?” (Aussems and Brink 294).

The Lahun town papyri offer a random sample of the handwriting from their time. This project foregrounds the distinctive archaeological combination of small size and large number of fragments. A digital approach to combined totality and fragmentariness might dictate a new quality of research and consciousness of results, true to the aim that “the digital representation of manuscripts determines scholarly work” (Vogeler xv). Specifically, digitisation delivers for analysis for the first time in practice a ‘totality’ of writing from Lahun 1800 BC constituted by the random sample preserved through site history and ground conditions. Computing capacity here shifts from auxiliary “computer-aided” to “computer-enabled palaeography”. The totalising horizon of the sheerly fragmentary calls for massively increased quantifiable precision, and stimulates more radical ambition from interdisciplinary history: to delineate for analysis an unprecedented profile of literacy in one time/place case-study—Egypt 1800 BC. Computer-enabled palaeography moves Egyptology on to a more open forum of cultural and historical studies.

By revalorising the quantitative, and theorising the qualitative aspect of the archaeological fragment, the project delivers one specific consequence of considerable importance for studies of ancient literacy: a renewed focus on the aspect of numeracy. Accountancy documents constitute by far the greater part of the Lahun manuscripts, from both temple and town. Yet narrative, in particular literary, attracts vastly more attention in Egyptology (Moreno Garcia). A project involving different sciences may return us to the starting-point of the writing, which is, in bulk, counting. Here the mark that lies on the border of, or even outside, script, remains under-researched for Egyptian manuscripts, but precisely this border is providing fruitful ground for digital palaeography initiatives (Hofmeister et al. 278). Lahun check-marks are distinguished sometimes by form, but also by alignment and repetition; on fragment UC32130 a workforce name-list deploys check-marks for names, but also repeats the sign for “child” in check-mark style. In other instances, marks may operate differently: fragment UC32107A presents sequences of marks that occur on several other papyri, interpretable as multiplication device (Imhausen). Numeracy and literacy cannot be understood without the other; formal and quantitative analytical programmes bring out forcefully the interpenetration of these two dimensions of writing.

By providing new access and flexibility to the fixed signs on the papyrus-paper, the interdisciplinary approach contributes new resources in tackling some of the deepest

problems in history-writing. In practice we still remain silent on the gaps in our knowledge of ancient literacy, which we still tend to fill with cultural assumptions on ethnicity and class. The dual motif of Scribe with Priest has a notable history of abuse (Ferro) in European literatures, including academic writing, reifying the social relations of worlds that are ‘other’. In Eurocentric histories of north Africa and west Asia, the motif of a Scribe-Priest society continually reinscribes orientalist attitudes in disciplinary treatments of ancient literacy. In considerations of class, wittingly or otherwise, historians and archaeologists have tended to leave intact assumed absolute dividing-lines in part-literate societies—assumptions that may have most to do with reinforcing faultlines of their own societies. Archaeological provenanced fragments might alter discussions of ‘integral palaeography’ (Ciula 221) over the extent to which writers at particular times and places did or did not perceive themselves as a cohesive community.⁸ In such areas I anticipate repercussions from the ability to match signs to ‘hands’, and on this basis to begin to discuss the meaning as well as identities of ‘hand’. This opening may be the critical contribution from datasets of archaeological fragments, if the equation ‘hand’ = ‘a person’ remains the fundamental assumption required for the palaeographical operation (Stokes 317). Exponential acceleration of handwriting research in fragmented datasets could simultaneously re-focus attention from identifying individual writer or author(ity), to investigating of broader, more communal lives of those individuals, replacing the question ‘who/where is this writer’ with ‘how did writing live’ in a particular time/place (cf Stansbury 248).

Computing approaches begin not by replacing, but by enhancing and extending previous skills, and their application requires a balance in design of a developing research agenda. Palaeographers may refine localised questions manually, and bring insights of historical subject specialisms to bear on quantitative results. However, despite endorsement by critical writers such as Benjamin and Gramsci (Crehan 30–31), philology alone offers little defence against dehumanisation, after its role in constructing the inhuman power relations of the colonial regime. To defuse the philologist, a differently skilled, third party must at some point be invited, in order to retain the most human element in the story, the human body: the practised calligrapher seems best placed to assess automated analyses for “reconstructing the motion of the scribe’s pen on the page” (Stansbury 248). These three roles could create out of quantitative and qualitative results a new agenda for understanding major changes beyond individual or collective intention. At the same time, such multiple partnerships might be better placed to foster “the opening towards the society of the research work and its transparency” (Cartelli and Palma 133). The abrupt shift in writing style around 1850 BC belongs within a major transformation in material cultural history. Handwriting fits tangibly

⁸ This approach seeks to problematise productively each corner of the clearly articulated model individual-community-society, Cartelli and Palma 132.

into patterns of change, but without clear motivations or agendas yet identified. What can such a seemingly innocuous shift express or reflect? Will measuring it help us? The social historical expectations on the digital age run ambitiously high.

Bibliography

- Andrén, Anders. *Between Artefacts and Texts: Historical Archaeology in Global Perspective*. New York: Plenum Press, 1998.
- Assmann, Jan. "Gebrauch und Gedächtnis. Die zwei Kulturen des pharaonischen Ägypten." *Kultur als Lebenswelt und Monument*. Eds. Aleida Assmann and Dieter Harth. Frankfurt a.M.: Fischer, 1991. 135–152.
- Aussems, Mark and Axel Brink. "Digital Palaeography". *KPDZ* 1. 293–308.
- Bosson, Nathalie and Sydney Aufrère. *Egyptes... l'égyptien et le copte*. Lattes: Musée de Lattes, 1999.
- Butler, Beverley. *Return to Alexandria. An Ethnography of Cultural Heritage Revivalism and Museum Memory*. Walnut Creek: Left Coast Press, 2007.
- Callinicos, Alex. *The Resources of Critique*. Cambridge and Malden: Polity, 2006.
- Cartelli, Antonio and Marco Palama. "Digistylus – an Online Information System for Palaeography Teaching and Research." *KPDZ* 1. 123–134.
- Ciula, Arianna. "The Palaeographical Method under the Light of a Digital Approach." *KPDZ* 1. 219–235.
- Coleman, Joyce. *Public Reading and the Reading Public in Late Medieval England and France*. Cambridge: Cambridge University Press, 1996.
- Collier, Mark. "Lots I and II from Lahun." *Archaism and Innovation. Studies in the Culture of Middle Kingdom Egypt*. Eds. David Silverman, William Kelly Simpson and Josef Wegner. New Haven: Yale University Press, 2009. 205–259.
- Collier, Mark and Stephen Quirke. *The UCL Lahun Papyri: Letters*. Oxford: British Archaeological Reports, 2002.
- Collier, Mark and Stephen Quirke. *The UCL Lahun Papyri: Religious, Literary, Medical, Legal*. Oxford: British Archaeological Reports, 2004.
- Collier, Mark and Stephen Quirke. *The UCL Lahun Papyri: Accounts*. Oxford: British Archaeological Reports, 2006.
- Crehan, Kate. *Gramsci, Culture and Anthropology*. London: Pluto, 2002.
- Delange, Elisabeth et al. "Apparition de l'encre metallogallique en Égypte à partir de la Collection de Papyrus du Louvre." *Revue d'Égyptologie* 41 (1990): 213–217.
- Dévaud, Eugène. *L'âge des papyrus égyptiens hiératiques d'après les graphies de certains mots: de la XII^{me} Dynastie à la fin de la XVIII^e Dynastie*. Paris, 1924.
- Drower, Margaret. *Letters from the Desert: the Correspondence of Flinders and Hilda Petrie*. Oxford: Archaeopress, 2004.
- Ferro, Marc. *Comment on raconte l'histoire aux enfants: à travers le monde*. Paris: Payot, 1981.
- Fischer, Henry George. *L'écriture et l'art de l'Égypte ancienne*. Paris: Presses Universitaires France, 1986.

- Foucault, Michel. *L'archéologie du savoir*. Paris: Gallimard, 1969.
- Frey, Rosa A. and James E. Knudstad. "The Re-examination of Selected Architectural Remains at El-Lahun." *Journal of the Society for the Study of Egyptian Antiquities* 35 (2008). 23–82.
- Grajetzki, Wolfram. *Lahun: Pictures of the Site Today*. London: University College London, 2002. <<http://www.digitalegypt.ucl.ac.uk/lahun/today.html>>.
- Gran, Peter. "Reply to review by F. de Jong of Islamic Roots of Capitalism: Egypt 1760–1840". *International Journal of Middle East Studies* 14 (1982): 399f.
- Gunn, Battiscombe. "The Name of the Pyramid-Town of Sesostris II." *Journal of Egyptian Archaeology* 31 (1945): 106–107.
- Helck, Wolfgang. *Die Lehre des Dw3-Htj*. Wiesbaden: Harrassowitz, 1970.
- Hofmeister, Wernfried, Andrea Hofmeister-Winter and Georg Thallinger. "Forschung am Rande des paläographischen Zweifels. Die EDV-basierte Erfassung individueller Schriftzüge im Projekt DAMaS." *KPDZ* 1. 261–292.
- Horváth, Zoltán. [2009a.] "Temple(s) and Town at el-Lahun. A Study of Ancient Toponyms in the el-Lahun Papyri." *Archaism and Innovation. Studies in the culture of Middle Kingdom Egypt*. Ed. Peter Der Manuelian. Boston: Museum of Fine Arts, 2009. 171–203.
- Horváth, Zoltán. [2009b.] El-Lahun Survey Project. [Website of the Office of the Hungarian Cultural Counsellor Cairo – Hungarian Embassy.] <http://www.magyarintezet.hu/index2.jsp?HomeID=7&lang=ENG&std_func=MIS&id=30717&high_art=false&page=>>.
- Imhausen, Annette. "UC32107A verso: a Mathematical Exercise?" [Commentary.] Collier and Quirke 2006: 288–291.
- Ives, Peter. *Gramsci's Politics of Language: Engaging the Bakhtin Circle and the Frankfurt School*. Toronto: University of Toronto Press, 2004.
- Jameson, Fredric. "How not to Historicise Theory." *Critical Inquiry* 34 (2008): 563–582.
- Janssen, Jac. "On Style in Egyptian Handwriting." *Journal of Egyptian Archaeology* 73 (1987): 161–167.
- Kaplony-Heckel, Ursula. *Verzeichnis der orientalischen Handschriften in Deutschland 19.1. Ägyptische Handschriften. Teil 1*. Wiesbaden: Harrassowitz, 1971.
- Kaplony-Heckel, Ursula. "Kahun-Papyri." *Lexikon der Ägyptologie* III. Ed. Wolfgang Helck. Wiesbaden: Harrassowitz (1980), cols. 292–293.
- KPDZ 1: *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Eds. Malte Rehbein, Patrick Sahle and Torsten Schaßan. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. Online: <<urn:nbn:de:hbz:38-29393>>, <<http://kups.ub.uni-koeln.de/volltexte/2009/2939/>>.
- Luft, Ulrich. *Die Chronologische Fixierung des ägyptischen Mittleren Reiches nach dem Tempelarchiv von Illahun*. Vienna: Österreichische Akademie der Wissenschaften, 1992.
- Möller, Georg. *Hieratische Paläographie I–III*. Leipzig: J.C. Hinrichs, 1909–1936. Online at <<http://www.egyptology.ru/lang.htm#Moeller>>.
- Moreno Garcia and Juan Carlos. "From Dracula to Rostovtzeff, or the Misadventures of Economic History in Early Egyptology." *Das Ereignis. Geschichtsschreibung zwischen Vorfall und Befund*. Ed. Martin Fitzenreiter. London: Gold House Publications, 2009. 175–198.
- Müller, Matthias. Fragment einer Beschwörung aus dem Mittleren Reich. *Göttinger Miszellen* 216 (2007): 51–54.

- Petrie, William Matthew Flinders. *Kahun, Gurob, Hawara*. London: Kegan Paul, Trench, Trübner and Co., 1890.
- Petrie, William Matthew Flinders. *Illahun, Kahun, Gurob*. London: Kegan Paul, Trench, Trübner and Co., 1891.
- Posener-Kriéger, Paule. *Les Archives du Temple Funéraire de Néferirkarê-Kakaï (les papyrus d'Abousir)*. Cairo: Institut Français d'Archéologie Orientale du Caire, 1976.
- Quaegebeur, Jan. "Books of Thoth Belonging to Owners of Portraits? On Dating Late Hieratic Funerary Papyri." *Portraits and Masks. Burial Customs in Roman Egypt*. Ed. Morris Bierbrier. London: British Museum Press, 1997. 72–77, pl. 35.
- Quirke, Stephen. *Hidden Hands. Egyptian Workforces in Petrie Excavation Archives, 1880–1924*. London: Duckworth, 2010.
- Quirke, Stephen. "Labour at Lahun." *The Archaeology and Art of Ancient Egypt*. Eds. Zahi Hawass and Janet Richards. Cairo: American University in Cairo Press, 2007.
- Recensement général. Recensement général de l'Égypte 1er juin 1897 – 1er Moharrem 1315*. Cairo 1897, Section VI: Instruction [without information on age in literacy].
- Roccati, Alessandro. Scrittura e testo nell'antico Egitto. *Scrittura e civiltà* 15 (1991). 21–31.
- Shubert, Stephen. "Does She or Doesn't She? Female Literacy in Ancient Egypt." *Proceedings of the Near and Middle Eastern Civilizations Graduate Students' Annual Symposia 1998–2000*. Toronto: Benben Publications, 2001. 55–76.
- Stansbury, Mark. "The Computer and the Classification of Script." *KPDZ* 1. 237–249.
- Stokes, Peter. "Computer-aided Palaeography, Present and Future." *KPDZ* 1. 309–338.
- Tait, John and Bridget Leach. "Papyrus." *Ancient Egyptian Materials and Technology*. Eds. Ian Shaw and Paul Nicholson. Cambridge: Cambridge University Press, 2000. 227–253.
- Valbelle, Dominique. "Deir el-Medineh." *Lexikon der Ägyptologie* I. Ed. Wolfgang Helck. Wiesbaden: Harrassowitz (1975), cols. 1028–1034.
- Vogeler, Georg. "Einleitung. Der Computer und die Handschriften. Zwischen digitaler Reproduktion und maschinengestützter Forschung." *KPDZ* 1. xv–xxiv.

Recognizing Degraded Handwritten Characters

Markus Diem, Robert Sablatnig, Melanie Gau, Heinz Miklas

Abstract

In this paper, Slavonic manuscripts from the 11th century written in Glagolitic script are investigated. State-of-the-art *optical character recognition* methods produce poor results for degraded handwritten document images. This is largely due to a lack of suitable results from basic pre-processing steps such as binarization and image segmentation. Therefore, a new, binarization-free approach will be presented that is independent of pre-processing deficiencies. It additionally incorporates local information in order to recognize also fragmented or faded characters. The proposed algorithm consists of two steps: character classification and character localization. Firstly *scale invariant feature transform* features are extracted and classified using *support vector machines*. On this basis interest points are clustered according to their spatial information. Then, characters are localized and eventually recognized by a weighted voting scheme of pre-classified local descriptors. Preliminary results show that the proposed system can handle highly degraded manuscript images with background noise, e.g. stains, tears, and faded characters.

Zusammenfassung

In diesem Beitrag werden slawische Manuskripte aus dem 11. Jahrhundert analysiert. Herkömmliche *Optical Character Recognition* (OCR) Systeme erzielen schlechte Resultate auf den beschädigten glagolitischen Schriften, da eine korrekte Buchstabensegmentierung nicht möglich ist. Deshalb wird ein segmentierungsfreies OCR-System vorgestellt, welches keiner Vorverarbeitungsschritte bedarf. Da die Klassifikation auf lokaler Information beruht, ist es möglich auch verblasste Buchstaben bzw. Buchstabenfragmente richtig zu erkennen. Das System besteht aus zwei grundlegenden Methoden: Buchstaben-Klassifizierung und Buchstaben-Lokalisierung. Die Klassifizierung basiert auf lokalen, größeninvarianten Merkmalen, die mit Hilfe von *Support Vector Machines* klassifiziert werden. Nach diesem Schritt existieren mehrere gekennzeichnete Merkmalsvektoren pro Buchstabe. Diese werden im zweiten Schritt durch ein *Clustering* Verfahren zusammengefasst, so dass jedem Buchstaben ein finales Klassenetikett zugewiesen werden kann. Die Ergebnisse zeigen, dass auch beschädigte Dokumente mit diesem System automatisch erfasst werden können.

1. Introduction

In the digital age Optical Character Recognition (OCR) has been successfully established for automated document analysis of standardized, typeset text. The automatic decipherment of handwriting, however, still poses difficulties for modern and ancient documents likewise. Even more so this applies to damaged or degraded material, the script of which is no longer readable straightforwardly.

In 2007 the interdisciplinary project “Critical Edition of the New Sinaitic Glagolitic Euchology (Sacramentary) Fragments with the Aid of Modern Technologies” of philologists (University of Vienna), computer scientists (image processing group CLV, Vienna University of Technology) and material chemists (Vienna Academy of Fine Arts) was launched to analyse and edit two – later three – valuable Slavonic manuscripts, parchment codices of the Old Church Slavonic canon dating from the 11th century: the so-called *Missale Sinaiticum* (Sin. slav. 5/N), a sacramentary fragment consisting of approximately 70 folia written by one main and two minor hands, and a 28-folia fragment, the *Euchologium Sinaiticum pars nova* (Sin. slav. 1/N), part of the famous Sinaitic Euchology discovered in the 19th century. They are written in Glagolitic script, which was created in 862/3 by St. Constantine-Cyril for his mission in Great Moravia¹, and belong to the complex of new findings made in St. Catherine’s Monastery on Mt. Sinai in 1975. Both codices are of a small format of approximately 140x100 mm and are decorated with colour initials and headline highlighting in yellow and green. Unfortunately, especially the Missal shows extensive damages like faded ink, blurring of the ink, staining due to mould or humidity, degradation of the parchment, e.g. chipping, fragmentation and contortion of folia, and the rare phenomenon of chemical conversion of black into white ink. The manuscripts partly contain palimpsest (re-written) folia (for further reference see Miklas 2000).

In the course of the project we have explored several computational approaches to ease codicological and palaeographic investigations, such as layout analysis, semi-automatic character segmentation and feature extraction using graphetic distinctive features – as opposed to this approach –, initial detection and automated puzzling (cf. Kleber and Sablatnig 2009). Furthermore, we have investigated in particular on the description, decipherment and reconstruction of (latent) texts of the manuscripts in question with methods of multi-spectral imaging, image binarisation, document image analysis and image enhancement (Lettnner et al. forthcoming; Miklas et al. 2008). With the combined approaches the readability of the *Missale Sinaiticum* could be enhanced up to 51% (Miklas et al. forthcoming).

Due to the heavily degraded condition of our objects common OCR methods based on robust binarization algorithms did not show satisfying results. Consequently our

¹ Designed for liturgical use, the original Glagolica comprised 36 letters functioning also as numerals and fraught with various aspects of theological symbolism (Miklas 2003)

new system for OCR and character supplementation is based on an entirely binarization free method. It performs three major steps, which will be discussed in the following sections: First, the local features of a whole manuscript page are computed and classified by means of Support Vector Machines (SVM). Then, a clustering of interest points according to their spatial coordinates and scale enables the localization of characters. This results in probabilities for character classes that are the basis of a voting scheme for character labels. Preliminary results show that the proposed system can handle images of severely damaged manuscripts with low contrast of text and background and fragmented characters.

2. Related Work – Technical Overview

In this section, state-of-the-art OCR systems for degraded documents are presented. It is not intended to give a comprehensive overview (which was already done by Plamondon and Srihari; Vinciarelli), but to describe current developments in the recognition of historic manuscripts. To our knowledge no OCR system has been proposed that can extract features from gray-scale or color images. Current OCR systems have three basic steps in common: First, a document pre-processing is performed. There, the document's skew is estimated, the text layout extracted, and the document image binarized. Subsequently, binary features are extracted and classified by means of a Neural Network (NN) or a SVM. The approaches differ according to the investigated data. Generally, two data sets can be distinguished: For cursive handwritten documents a word based approach is chosen, for non-cursive, usually ancient manuscripts, a character based approach.

Cursive handwritten documents: Lavrenko et al. directly recognize words from documents of the George Washington collection. Their technique was later improved by Rath and Manmatha, who added compensation to non-linear variations present in manuscripts. Another word recognition system is proposed by Frinken and Bunke. They compute statistical moments from sliding windows that are applied to normalized word images. Hofmeister et al. compare sample word-forms (*templates*) for scribe identification.

Historical, non-cursive documents: In contrast to word recognition methods, Alirezaee et al. developed a character recognition system for medieval Persian manuscripts. They extract statistical features from previously binarized document images. Arrivault et al. propose a combined statistical and structural character recognition approach for ancient Greek and Egyptian documents. Here, structural features such as attributed graphs are computed and classified for characters rejected during the (preceding) classification of statistical features. Another approach concerning historical Greek documents was published by Vamvakas et al. that calculates zone

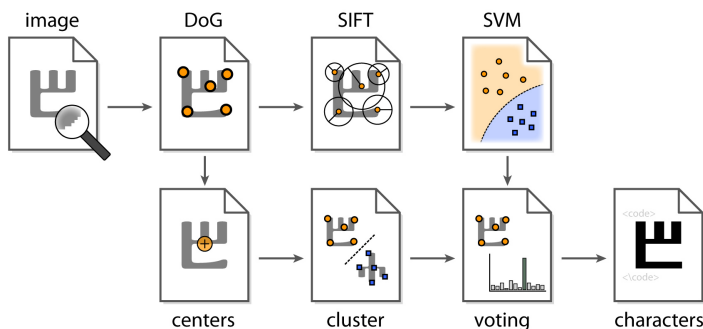


Figure 1. The proposed system consists of two task-levels: classification (upper row) and character localization (lower row).

features and character profile features on the binarized image segmented individual characters. In 2007 Ntzios et al. developed a so-called segmentation-free character recognition system applicable to the same document type. It extracts geometrical features from binarized images in combination with a watershed-like algorithm that fills cavities. A decision tree is used for the character classification.

The BIT-Alpha company (Tomasi and Tomasi 2009) presents a combined approach considering both word and character detection methods.

However, none of these methods gives positive results with faded and damaged manuscripts.

3. Methodology

Contrary to the methods introduced in the previous section, the proposed system here has a fundamentally new architecture, which is designed to compensate the drawbacks that arise when dealing with ancient manuscripts. Instead of applying a binarization in order to compute features, they are directly extracted from the gray-scale image.

According to its major tasks, the system is divided into two procedures: classification and localization (see Fig. 1). The fulfillment of both tasks is based upon the extraction of interest points. The interest point detector extracts blob-like regions at different scales of the manuscript image. Local descriptors robustly define the respective regions with respect to a certain set of image transformations. The descriptors are then classified by means of a multi-kernel SVM. Having classified all extracted image regions, each character consists of multiple pre-classified points. In order to assign one class label to each character present in an image, the interest points need to be clustered. *K*-means clustering groups the interest points according to the underlying characters. Finally,

a so-called interest point voting weights the class probabilities of all local descriptors belonging to the same cluster and assigns the final class label to every character.

3.1. Feature Extraction

As outlined, characters are detected and classified by means of interest points. These points are located at local extrema of the image's second derivative which is approximated via the Difference-of-Gaussians (DOG) function (Lowe). In other words, the interest points mark character attributes, such as junctions, endings, stroke borders, corners, and circles, and their respective size.

In order to mathematically describe the characters' attributes, local descriptors are computed at the locations of interest points. Intuitively, one could consider the gray-values of the image within a local grid (as they are the basic information which is observed by humans, too). However, this information is not robust against image transformations such as affine transformations (e.g. rotation, scale) or photometric changes (illumination, sensor noise). That is why local descriptors are computed at locations of interest points. The proposed system computes Scale-Invariant Feature Transform (SIFT) descriptors, which were first introduced by Lowe. These descriptors convert a local image grid into a 128-dimensional vector by means of gradients. Thus, the descriptors are robust against photometric changes. Additionally, SIFT descriptors are invariant with respect to rotation and scale. Considering the challenge of character recognition, it is desirable to recognize characters of different size or resolution. However, the descriptor's invariance to rotation leads to problems when recognizing characters. For instance, the Glagolitic Ѣ has the same topology as the Glagolitic Ѧ, rotated by 180°. If we consider descriptors that are invariant to rotation, the system cannot differentiate between these two characters. That is why the SIFT descriptors are not computed rotationally invariant, but robust against rotational changes.

Fig. 2 shows two Glagolitic characters with their corresponding interest points. Gray circles show the region of interest points which are denoted by white squares. The lines connecting the squares with the circles indicate the main orientation of each interest point. The histograms represent down-sampled local descriptors of the highlighted (black) interest point. Note that the local descriptors are the same if they are computed rotationally invariant (360°).

3.2. Classification

Having computed the local descriptors, character attributes – such as strokes, junctions or endings – are described by high-dimensional feature vectors. Assuming that the character attributes slightly change from one character class to another, characters can be recognized using only local information. In other words, there is no need for

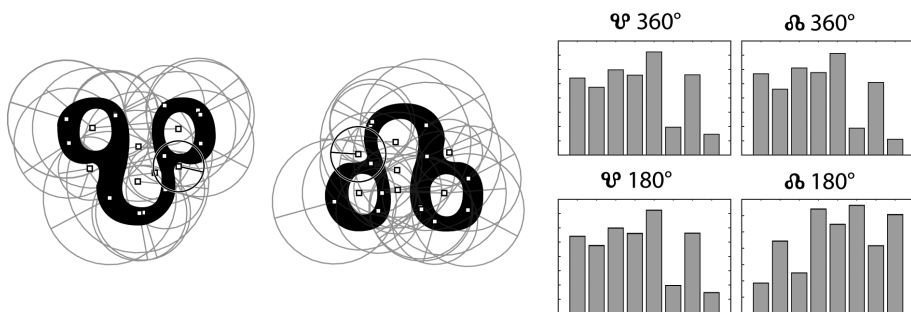


Figure 2. A Glagolitic \mathcal{U} and \mathcal{B} with their local descriptors (left). The down-sampled features computed rotationally invariant (right) and with rotational dependence up to 180° .

establishing a relationship between the local descriptors of one character in order to recognize the character. That is why they are directly classified in the proposed system. We use SVMs, which can be trained on classifying local descriptors – in our case character features – to assigned classes. SVMs benefit from the fact that they are based on statistical learning theory rather than error minimization. Thus, they achieve a good generalization – while still being flexible – even if a small training set is obtained. For the classification an SVM is trained using 20 manually tagged characters per character class.

In order to further improve the classification performance, one-against-all tests are performed. This means that one SVM is trained per character class. Each SVM decides whether a local descriptor belongs to its character class or not (e.g. \mathcal{U} , not \mathcal{U}). In addition to the class labels predicted, a probability is assigned by each classifier resulting in a probability histogram, i.e. the assignment of the probability of each interest point of belonging to a certain character class (cf. Fig. 3). Another advantage of one-against-all tests is the fact that the classifiers are not too sensitive to noise in the training data, as the criterion function is less complex when two classes are to be considered.

3.3. Character Localization

For traditional OCR engines, the characters or words are localized implicitly in the binarization step. If handwriting OCR engines are considered, an additional character segmentation step needs to be performed in order to detect concatenated characters. In contrast, the proposed system has no information about the positions of characters in a given image to the point of feature classification. Indeed, the positions of the classified features are known, but – as a feature does not necessarily represent a whole character – the position and size of the character is unknown. The character localization is based

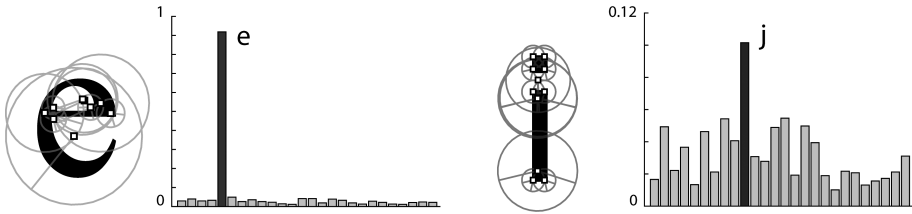


Figure 3. Probability histogram of two character clusters. A correct classification (left) and a false classification (right).

on clustering the interest points in the spatial domain, in our case applying the k -means clustering algorithm (see following paragraph). This approach benefits from the fact that degraded characters are detected with local descriptors, but not considered when the image is binarized. Thus, even degraded characters can be localized. Another advantage is the low computational complexity, since only the interest points are considered.

The k -means algorithm groups clusters of interest points and their centers. Each group should correspond to one character. But the k -means clustering cannot estimate the number of clusters k . To overcome this problem the scales of interest points are exploited. Each character produces a single local maximum in a certain scale level. When this information is extracted, the number k of the k -means can be estimated and at the same time initial cluster positions are obtained that improve convergence. Having clustered the interest points, each cluster consists of all interest points that belong to the same character.

3.4. Feature Voting

For the final character classification a voting scheme is applied. Therefore, all local descriptors of a cluster are considered. Each descriptor was previously classified. Hence, a probability histogram exists that indicates the class likelihood of each descriptor in the cluster. If these histograms are accumulated, the maximum bin indicates the most probable class label. Fig. 3 shows the final probability histogram of two degraded characters. Each histogram bin represents one of the previously trained character classes. The bin's height indicates each character's probability of belonging to the respective class. The left character is classified correctly, having a significantly high class probability. In contrast, the probability histogram of a false classification is given in Fig. 3 (right). There, three class probabilities are similarly high. If the histogram is indecisive, e.g. for character fragments, the alternative hypotheses could be further processed, e.g. by a dictionary.

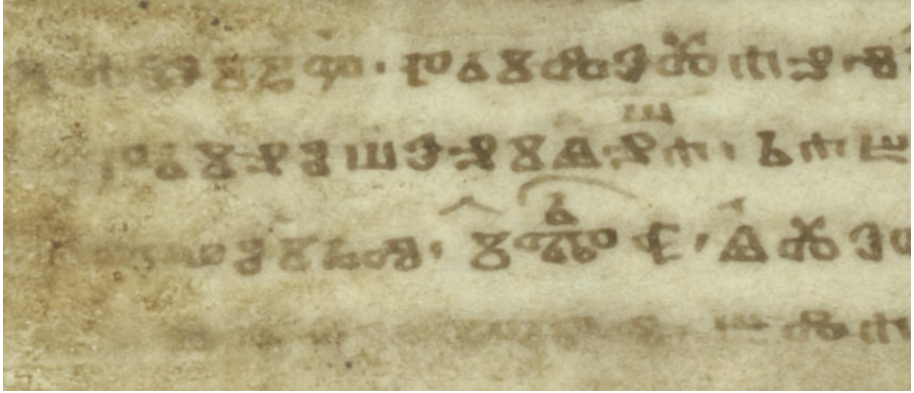


Figure 4. Random sample page of *Missale Sinaiticum*.

4. Results

In this section the evaluation of the proposed system is given. In order to evaluate the system, 15 pages containing 1055 characters are extracted from the *Missale Sinaiticum*. The pages were chosen randomly (cf. Fig. 4). They contain faded-out ink, degraded characters and background noise. For groundtruthing, each character was brushed with a gray-value that corresponds to its class index.

The evaluation is based on the values of True Positives (TP) (correctly located and correctly classified characters), False Positives (FP) (correctly located, but falsely classified characters) and False Negatives (FN) (characters which are not located). These values allow for computing the precision and recall. Thus, the precision indicates the percentage of correctly classified characters to those retrieved. Whereas the recall specifies the percentage of correctly classified characters to those present in an image. Mathematically, the former is defined as the sum of TP divided by the sum of retrieved values ($TP + FP$). The latter is the sum of TP divided by the total number of elements that exist ($TP + FP + FN$).

The aim of a classification task is to maximize both, the precision and the recall. Therefore the F score is introduced, which is a weighted average between the precision and the recall:

$$F\beta = \frac{(1 + \beta^2)p \cdot r}{\beta^2 p + r} \quad \leftrightarrow \quad F\beta = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2 FN + FP}$$

	F0.5-score	recall	precision	#
with clustering	0.772	0.673	0.832	1055
artificial clustering	0.804	0.748	0.837	1055

Table 1. System's recall, precision and F-score when the proposed system and artificial clustering is applied.

where r is the recall and p is the precision. The right equation expresses the F-score in terms of TP/FP . The β allows weighting the precision or the recall. Thus, if β is set to 0.5, the precision is weighted twice as much as the recall.

System Evaluation: In order to demonstrate the effect of the character localization, artificial clustering is implemented. This is based on the annotated ground truth where cluster centers are defined as the center-of-mass of each blob. As a constraint, only interest points being within a character blob are considered. Therefore, the character localization (clustering) does not introduce an error. Thus, the error introduced by clustering can be extracted. The system achieves an F0.5-score of 0.772 on the investigated dataset. If artificial clustering is applied, an F0.5-score of 0.805 is achieved. This directly draws the conclusion that the F-score is decreased by 0.033 because of the character localization. The test setup additionally shows that the character clustering has hardly any influence on the system's precision (difference: 0.005). In contrast, the proposed k -means decreases the recall rate by 0.075. This results from clustering errors which increase the FN rate as characters are not localized correctly.

Evaluation of Degraded Characters: By extracting single characters, it is possible to evaluate only the classification step illustrated in Fig. 1. Therefore two datasets are constructed that consist of single characters which were annotated and extracted from the Missal.

The first dataset (setA) consists of 10 classes having 10–12 samples (totally 107) which are well preserved. This dataset is a reference for the evaluation with degraded characters. The second dataset, which is referred to as setB, contains 25 character classes with approximately 9 characters per class (totally 198). Degraded or partially visible characters were extracted to construct this set. It is used to demonstrate the system's behavior when degraded characters need to be recognized.

Fig. 5 shows examples of both datasets. It can be seen that some characters such as \mathcal{O} , \mathcal{U} and \mathcal{V} are similar to each other. The degraded characters in the second row differ strongly from those of setA. They are hard to read for humans.

SetA is evaluated first in order to show the performance of the method on undistorted data. Therefore, 10 SVM kernels are trained using 10 samples per class. Then all 107 test characters are evaluated. The voting is the same as described in Section 3.4, except for

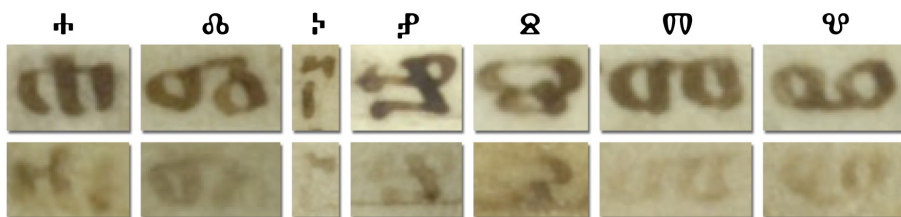


Figure 5. Examples of the datasets evaluated. The first row shows characters of setB, whereas the second row shows the same examples from the dataset containing degraded characters.

the fact that the clustering needs not to be performed. For the character classification an overall precision of 98.13% is achieved, which means that only 2 characters out of 107 are falsely predicted. Both confused characters consist of two circles and a connecting stroke (see Fig. 4 second and last column) which produce similar descriptors.

For a direct comparison of both datasets, the same ten classes were extracted from setB. Certainly the same classifier is used in both test setups. In contrast to setA the degraded characters in setB have a lower precision which is 78.89%. These numbers indicate that it is harder for the system to classify degraded characters. On the other hand the system can cope with uncertainty which arises from the fact that fewer descriptors are classified in this case.

In addition to the comparison of setA and setB, all 198 degraded characters were evaluated. Even though 25 different classes are predicted in this evaluation (+15 classes), the precision decreases slightly by 7.17%. Thus, the overall precision is 71.72% when descriptor voting is applied on degraded characters. The ratio of detected descriptors and those classified now is 26%, which means a decrease by 13% compared to the previous test on the same dataset with 10 classes. Since the performance decrease is lower than the complexity increase, the system proves to be capable for classifying degraded manuscripts.

5. Conclusion

This paper shows a new methodology for character recognition of ancient manuscripts. The approach, which is inspired by recent object recognition systems, exploits local descriptors directly extracted from gray-scale images. Multiple SVMs are used to classify the local descriptors. The character localization is based on clustering interest points previously extracted for the computation of local descriptors. A scale selection that adapts to the manuscript image observed allows for the cluster center initialization.

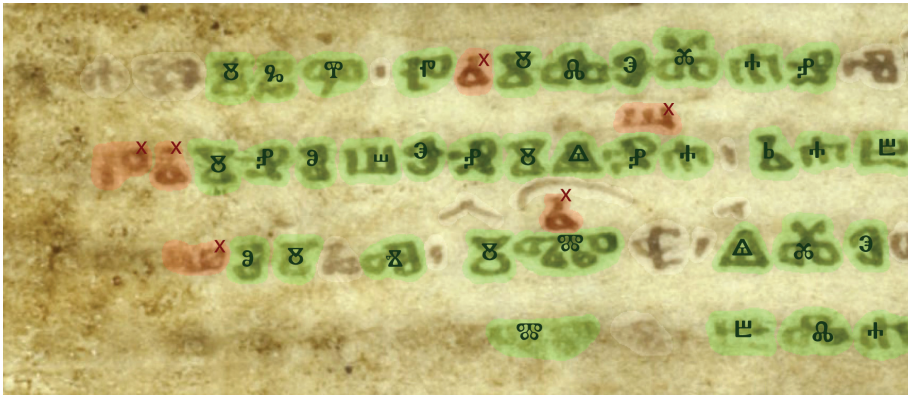


Figure 6. Sample page (cf. Fig.4) with results: Green blobs *TP*, red blobs *FP*.

The OCR system presented does not need any pre-processing of document images. In contrast to existing systems, a new architecture has been designed that focuses on images of degraded manuscripts. Since ancient manuscripts—much more often than modern—exhibit stains, faded-out ink and rippled support, they pose new challenges for OCR.

Bibliography

- Alirezaee, Shahpour, Hassan Aghaeinia, Karim Faez and Alireza S Fard. “An Efficient Feature Extraction Method for the Middle-Age Character Recognition.” *Advances in Intelligent Computing*. Berlin/Heidelberg: Springer, 2005. 998–1006.
- Arrivault, Denis, Noël Richard, Christine Fernandez-Maloigne and Philippe Bouyer. “Collaboration between Statistical and Structural Approaches for Old Handwritten Characters Recognition.” *Graph-Based Representations in Pattern Recognition*. Eds. Luc Brun and Mario Vento. Berlin/Heidelberg: Springer, 2005. 291–300.
- Frinken, Volkmar, and Horst Bunke. “Self-Training Strategies for Handwriting Word Recognition.” *Advances in Data Mining. Applications and Theoretical Aspects*. Eds. Petra Perner. Berlin/Heidelberg: Springer, 2009. 291–300.
- Frinken, Volkmar, Tim Peter, Andreas Fischer, and Horst Bunke. “Improved Handwriting Recognition by Combining Two Forms of Hidden Markov Models and a Recurrent Neural Network.” *Computer Analysis of Images and Patterns*. Eds. André Gagalowicz and Wilfried Philips. Berlin/Heidelberg: Springer, 2009. 189–196.
- Hofmeister, Wernfried, Andrea Hofmeister-Winter, and Georg Thallinger. “Forschung am Rande des paläologischen Zweifels: Die EDV-basierte Erfassung individueller Schriftzüge im Projekt DAMALS.” *KPDZ* 1. 261–92.

- Kleber, Florian, and Robert Sablatnig. "A Survey of Techniques for Document and Archaeology Artefact Reconstruction." *10th Int. Conf. on Document Analysis and Recognition (ICDAR)*. Barcelona, 2009. (CD publication).
- KPDZ 1: *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Eds. Malte Rehbein, Patrick Sahle and Torsten Schaßan. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. Online: <urn:nbn:de:hbz:38-29393>, <<http://kups.ub.uni-koeln.de/volltexte/2009/2939/>>.
- Lavrenko, Victor, Toni M. Rath, and R. Manmatha. "Holistic Word Recognition for Handwritten Historical Documents." *1st International Workshop on Document Image Analysis for Libraries (DIAL)*, 2004. 278–287.
- Lettner, Martin, Melanie Gau, Heinz Miklas and Robert Sablatnig. "Image Acquisition & Processing Routines for Damaged Manuscripts." *Digital Medievalist* (forthcoming).
- Lowe, David G. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision (IJCV)* 60 (2004): 91–110.
- Miklas, Heinz. "Die slavischen Schriften: Glagolica und Kyrillica." *Der Turmbau zu Babel. Ursprung und Vielfalt von Sprache und Schrift*. Ed. Wilfried Seipel. Wien: Kunsthistorisches Museum Wien & Skira, 2003. 243–49 (No. 3.5.16–26).
- Miklas, Heinz. "Zur editorischen Vorbereitung des sog. Missale Sinaiticum (Sin. Slav. 5/N)." *Glagolitica. Zum Ursprung der slavischen Schriftkultur*. Ed. Heinz Miklas. Wien: ÖAW, 2000. 117–29, XV–XVI.
- Miklas, Heinz et al. "St. Catherine's Monastery on Mount Sinai and the Balkan-Slavic Manuscript-Tradition." *Slovo. Towards a Digital Library of South Slavic Manuscripts*. Sofia, 2008. 13–36, 286.
- Miklas, Heinz, Melanie Gau, Martin Lettner, and Manfred Schreiner. "Editing the Sinaitic Glagolitic Inedita. State of the Art." *Codex Sinaiticus. Manuscripts in Modern Information Environment* (12.–13. Nov. 2009). St. Petersburg, forthcoming.
- Ntzios, Kostas, Basilios Gatos, Ioannis Pratikakis, Thomas Konidakis, and Stavros J Perantonis. "An Old Greek Handwritten Ocr System Based on an Efficient Segmentation-Free Approach." *International Journal on Document Analysis and Recognition* 9 (2007): 179–92.
- Plamondon, Réjean, and Sargur N. Srihari. "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2000): 63–84.
- Rath, Toni M., and Rudrapatna Manmatha. "Word Spotting for Historical Documents." *International Journal on Document Analysis and Recognition* 9 (2007): 139–52.
- Tomasi, Gilbert, and Roland Tomasi. "Approche informatique du document manuscrit." *KPDZ* 1. 197–218.
- Vamvakas, Georgios, Basilios Gatos, Nikolaos Stamatopoulos, and Stavros J Perantonis. "A Complete Optical Character Recognition Methodology for Historical Documents." *Document Analysis Systems (DAS)* 1 (2008): 525–32.
- Vinciarelli, Alessandro. "A Survey on Off-Line Cursive Word Recognition." *Pattern Recognition* 35, no. 7 (2002): 1433–46.

Finding What You Need, and Knowing What You Can Find: Digital Tools for Palaeographers in Musicology and Beyond

Julia M. Craig-McFeely

Abstract

This chapter examines three projects that provide musicologists with a range of resources for managing and exploring their materials: DIAMM (Digital Image Archive of Medieval Music), CMME (Computerized Mensural Music Editing) and the software Gamera. Since 1998, DIAMM has been enhancing research of scholars worldwide by providing them with the best possible quality of digital images. In some cases these images are now the only access that scholars are permitted, since the original documents are lost or considered too fragile for further handling. For many sources, however, simply creating a very high-resolution image is not enough: sources are often damaged by age, misuse (usually Medieval ‘vandalism’), or poor conservation. To deal with damaged materials the project has developed methods of digital restoration using mainstream commercial software, which has revealed lost data in a wide variety of sources. The project also uses light sources ranging from ultraviolet to infrared in order to obtain better readings of erasures or material lost by heat or water damage. The ethics of digital restoration are discussed, as well as the concerns of the document holders. CMME and a database of musical sources and editions, provides scholars with a tool for making fluid editions and diplomatic transcriptions: without the need for a single fixed visual form on a printed page, a computerized edition system can utilize one editor’s transcription to create any number of visual forms and variant versions. Gamera, a toolkit for building document image recognition systems created by Ichiro Fujinaga is a broad recognition engine that grew out of music recognition, which can be adapted and developed to perform a number of tasks on both music and non-musical materials. Its application to several projects is discussed.

Zusammenfassung

Dieser Beitrag stellt drei Projekte vor, die der musikwissenschaftlichen Forschung bei der Erschließung ihres Quellenmaterials dienlich sind: DIAMM (Digital Image Archive of Medieval Music), CMME (Computerized Mensural Music Editing) und das Programm Gamera. DIAMM verbreitet seit 1998 digitale Abbildungen von Handschriften in

bestmöglicher Qualität und macht sie der Forschung auf diese Weise weltweit zugänglich. In manchen Fällen bieten diese Bilder mittlerweile den einzigen Zugriff auf die Quelle, weil Originale verloren gegangen oder stark beschädigt sind. Bei vielen Handschriften reicht eine reine Digitalisierung in hoher Auflösung allerdings nicht aus: Sie sind auf Grund ihres Alters, einer nicht sachgerechten Behandlung oder schlechter Lagerung beschädigt. Für den Umgang mit beschädigtem Material hat das Projekt mit handelsüblicher Software Methoden einer digitalen Restaurierung entwickelt, die verlorengegangene Daten vielfältiger Materialien wieder sichtbar machen konnten. Darüberhinaus verwendet das Projekt ultraviolette wie infrarote Lichtquellen, um eine bessere Lesbarkeit bei Rasuren, Hitze- oder Wasserschäden zu gewährleisten. Der Beitrag diskutiert sowohl die Bedingungen digitaler Restaurierung als auch die Interessen der bewahrenden Institutionen. Mit CMME und einer daran angeschlossenen Datenbank musikgeschichtlicher Quellen und Editionen steht den Forschenden ein Hilfsmittel bereit, um dynamische Editionen und diplomatische Transkriptionen anzufertigen: Im Gegensatz zur starren Form einer bedruckten Seite kann ein digitales Editionsprogramm Transkriptionen für eine Reihe von unterschiedlichen Auszügen und Versionen weiterverarbeiten. Gamera schliesslich ist ein Bausatz für eine automatisierte Bilderkennung, der auf der Basis einer Software für Musikererkennung entwickelt wurde und auf eine Reihe von musikwissenschaftlichen und anderen Quellenmaterial mit jeweils unterschiedlicher Zielsetzung angewendet werden kann. Anwendung und Nutzen in verschiedenen Forschungsprojekten werden diskutiert.

1. DIAMM

Although today many libraries have embraced the idea of creating online access to their collections, there are very few (if any) subject-specific collections that cross borders of collection and country. The Digital Image Archive of Medieval Music (DIAMM) is a medieval musicology project that has managed to do this. Established in 1998 to collect images of endangered fragments and sources that are difficult to access, it has grown to embrace all major sources of medieval polyphony worldwide, and is gradually creating images of all these sources and adding them to the online collection. DIAMM undertakes the rather unglamorous job of providing a mass of research materials to the research community (and increasingly to a much wider public) in a digestible and useful way.

At the time of writing some 3136 manuscript sources of medieval polyphony (i.e. excluding chant) dating from before 1550 are known to musicologists. The number changes frequently as new sources (usually only fragments) are notified at a rate of several a year, particularly as archives in Spain and Italy are better explored. Polyphonic music from the period approx 800–1550 survives in a number of beautifully preserved

large, complete manuscripts ranging in extent from around 30 to 800 pages, and in size from about 12 x 15 cm up to books nearly a metre in height (designed so that a whole choir could sing from a single book simultaneously) from the Netherlands scriptorium of Petrus Alamire. Fragments add an enormous range to this corpus, equally varying in size from full sheets of huge choirbooks down to snippets the size of postage stamps, and ranging in extent up to large collections of as many as 50 leaves from larger books that have been dismembered but reunited more recently, either virtually or in the real world.

As with many medieval sources, there is an untapped wealth of fragments still to be discovered in the form of unremarked and uncatalogued endpapers, or boxes of bits and pieces that have not been thoroughly researched. Most existing library catalogues, prepared in the 19th or early 20th centuries, take no account of endpapers or binding fragments, and this neglect was reflected in the practices of rebinding well into the 20th century, when endpapers or paste-downs were routinely discarded when books were rebound. Some of the richest collections of fragments from bindings are to be found in libraries that did not enjoy a period of plenty in the Victorian era, when so many libraries embarked on wholesale rebinding of their collections.

The corpus is a realistic one to try to assemble under one umbrella: in comparison to other literatures the surviving witnesses are relatively few, but even with a controlled repertory the simple cost of creating images of all the sources and the time required to do so means that we are a long way short of completing the collection. Funding simply for creating and maintaining resources of this sort is difficult to come by, and the image collection presently grows only when a research project that requires images of a specific source is able to obtain funding to pay for digitization. We are also recipients of image donations from projects or institutions that have images but are unable (at present) to deliver them because of the cost of creating a delivery mechanism.

Central to the mission of DIAMM is to keep the resource free. Many online repositories of images are only available to those with personal or institutional subscriptions, cutting out the casual visitor, independent scholar, or interested amateur, all of whom have as much right to see the content as anyone else. Musicology is a small subject, and persuading a library with already limited resources to spend part of its budget on a subscription that might only benefit a very small percentage of its users is understandably difficult. Keeping access free has allowed users from all disciplines and from all walks of life to study, use, explore or simply enjoy this aspect of European cultural heritage.

Traditionally, research in this field has centred on the study of repertories preserved in large complete sources, often with a scholar dedicating their life's work to one particular manuscript. Access to other sources would depend on the scholar's ability to travel or to purchase photos or microfilms of inconsistent and often-dubious quality. Thus research

in the field has tended to become the province only of senior academics, with their detailed research limited to the times when they could physically visit a manuscript.

When DIAMM began digitizing documents we had no model on which to base a projection of how the creation and online delivery of images of rare and valuable documents would affect the access to them in the real world. Our initial assumption was that fewer scholars would need to visit the documents, and thus digitization could serve a preservation function since the documents would not be consulted in person as much as they were before the images were available. In at least two cases, these digital images are the only form in which the manuscript may now be consulted. We were certainly correct that a huge percentage of the work that had previously required physical contact with the manuscripts no longer needed to be done on site. However the explosion in internet usage from the turn of the century meant that manuscripts that had remained in relative obscurity for decades (even centuries) were now being brought to much wider public notice, and the new interest this aroused meant that a far broader public were interested in them and therefore wished to visit them.

This was not necessarily bad: for many libraries and archives their funding relies on demonstrable public access to their collections. Website hit-counter or visit analysis engines do not really provide the right sort of data for user analysis: many website visits are accidental or speculative, and the type of user and usage is difficult to pinpoint. Physical bodies walking through doors is still a much easier way to demonstrate usage of a physical resource, so the traffic to archives arising from their online exposure was not unwelcome.

A further concern in making images available online was that the sales of printed facsimiles might suffer. Again this proved unfounded: the British Library found that when they made images available online of manuscripts that had been published in facsimile, sales of the facsimiles increase.

1.1. Changes in Scholarship

The first major change to scholarship that digital imaging introduced was that scholars no longer had to work in black and white because of the cost constraints of colour photography. Colour imaging costs no more than black and white in the digital medium (although there are storage costs, since colour images are typically about three times the size of grayscale ones). The second change is that the ability of projects such as DIAMM both to create high-quality images and deliver them online in a one-stop free resource means that scholarship is no longer limited to the insular study of a single manuscript (or group of manuscripts), nor to those who can afford to buy the surrogates necessary for a broad grasp of the repertory as a whole, but can now take on a far more holistic attitude and approach.

The idea that all manuscripts can be available either on one website, or somewhere on the internet so that they can be accessed from a workstation anywhere in the world will not be new to readers, nor will the rude truth that probably less than 5% of all manuscript holdings are available online, and a good number of these require a paid subscription to view them. At the time of writing DIAMM delivers less than 20% of the known sources of polyphony in our designated field (although that number is rising), the limitation being due firstly to the fact that not all libraries will allow their images to appear online (though a surprising number embrace the facility provided by DIAMM wholeheartedly, since it costs them nothing), and secondly to the simple cost of buying or creating the images to add to the resource. The project has developed an enviable reputation in quality imaging and adds monthly to the online content thanks to direct donation of images, outside funding for imaging of specific projects or collections, and collaborations with, and consultancy for, other projects in which images created for them by DIAMM are also delivered through the DIAMM website.

An unanticipated benefit of digital imaging at high resolution has been that in many cases the detail shown in these images surpasses what is discernible to the naked eye, meaning that whereas previously a surrogate (photo, microfilm etc) was a poor alternative to seeing the original, an excellent digital image could now be considered better than the original, and in the case of several important endangered manuscripts DIAMM images are considered by their owners good enough to warrant withdrawal of the original source from public access.

What this means for palaeographers is a vast new repertory of fine-grained detail that was never—or only rarely—available to study before. The examination on-screen of this level of detail has allowed scholars to see erasures that are not visible to the naked eye, to examine ink density that shows where a scribe lifted his pen and dipped, supplementing the evidence behind scribal concordance, to examine the direction of hair follicles on skin surfaces and their groupings to determine whether two leaves were cut from the same skin, and to highlight or adjust particular areas of a leaf digitally to restore or improve legibility.

1.2. Dealing with Colour

Unfortunately as exemplars of accurate colour reproduction digital images cannot be taken and studied at ‘face value’ in the same way as analog surrogates, particularly when that examination is only undertaken on a computer screen. If study of the image only requires the ability to read the text, then a high-resolution colour digital image will provide detail and focus (assuming it is properly taken) well beyond what a hard copy can provide, even in colour. However if colour recognition is crucial for research then for some, the potential of the digital image is far from being fully realized, while for others the inexperienced use of digital images is misleading. With analog surrogates the

control of colour and resolution is in the hands of the creator so that what is delivered to the end-user is controlled at the point of creation. In the digital medium, even if the image is carefully colour calibrated and produced at the highest resolution, the creator has no control over what happens to that image after it leaves their desktop.

Many computer users, particularly older users with deteriorating eyesight, set their screens to low resolution so that everything appears larger. Unfortunately this also means that everything on the screen is slightly fuzzy. Less common now (fortunately) is to find a monitor set to 256 colours, but it is sometimes deliberately done (particularly on older machines) to speed up screen redraw. This causes the loss of fine colour gradations, which break up into blocks. Screen colour is a long-standing problem: even matching screens calibrated with a spider and professional software often will not match in colour cast, and this level of calibration is virtually unknown outside imaging departments and professional print studios. It is also still surprisingly uncommon in digital imaging studios.

The digital facsimile therefore has not really come into its own—and possibly may never do so because of the huge variation in monitor type, quality and colour-cast, and the even greater variation in the needs and colour-sensitivity of users. The DIAMM website delivers images with a request to new users to perform a basic calibration on their screen, but we know that no matter how ‘perfect’ our images are at the point of capture, we can never deliver perfect colour to a researcher’s desktop unless we deliver in print. It is interesting that even in the digital age, when books and images are very easily and cheaply available in digital form, the majority of people still prefer to carry a paperback novel, and still enjoy the look and feel of a beautifully reproduced facsimile accompanied by a scholarly study, preferring this to digital access.

Digital images cannot simply be used in the same way as conventional photographic surrogates. Some years ago a scholar wrote that the new technology would allow comparison of ink colours (among other things) leading to new discoveries about paper preparation, inks and scribal concordances. What that person did not understand was that colour is one of the most devastatingly misleading fields in the digital imaging world.

We can take our camera and lighting equipment all over the world and take pictures that we believe are ‘the same’ in quality and colour balancing, but without an understanding of colour profiles and gray balancing, lay users simply cannot compare any two pictures effectively. Most users simply cannot trust the colour that they see on their screens, nor can they effectively compare the colour of two images taken with different equipment or by different photographers even using the same equipment unless they know what they are looking for. Monitors come out of their boxes miscalibrated, with a variety of colour casts, so that even if the service delivering the images to the screen knows what it is doing in colour profiling, there is no guarantee what the scholar

will see. Colour-blindness issues aside,¹ at least when we produce a colour facsimile we know that every person who uses it will be given material with the same colours. When we produce a digital facsimile or online image resource we have no control over what the user sees, we cannot even ensure they have their screen set to a sensible resolution or bit depth, so the images may appear blurry or in only 256 colours rather than millions.

The digital age that has been hailed as a great step forward for scholarship is still a minefield in imaging that is poorly understood, and one of the biggest pitfalls is the fact that those accessing the resources do not understand the potential deficiencies or shortcomings of the materials they are now using.²

Even assuming a ‘perfect’ digital image is delivered to the end user, there are problems in understanding that data. The problems however are massively compounded by image suppliers who create poor quality surrogates: images that were out of focus at the point of capture and have been ‘sharpened’; under- or over-exposed images that have been level-adjusted; images of faded, damaged or dirty sources that have been ‘improved’ by the image manager so that they look clearer; batches of images that did not look the same as the manuscript when checked on an uncalibrated monitor, and have been colour-shifted by the operator to ‘look right’ on that monitor; imaging processes that apply lossy compression formats such as JPEG during the workflow, thus losing detail and colour refinement. In order to look at a digital image now scholars need to know that what they are seeing may not be an accurate representation of the original, and that accuracy may have been compromised in ways that they are not qualified to identify.

There are no easy answers to closing the gap between supply and understanding of the product by the end user. We can identify the disease, but we cannot cure it, nor can we ensure that image-suppliers operate to a baseline of quality that ensures the end-user is not being supplied with poor quality information, thus misleading them further: a huge number of major research institutions worldwide send images of spectacularly poor quality from their digital studios. Institutions employ outside suppliers to do their imaging work, and then rely on them for quality assessment of their own output. Several organizations (including DIAMM) have set down their own quality guidelines, most of which agree, but there is no protocol that ensures those guidelines will be followed, even if they are cited when ordering images.³

Deficiencies aside, the revolution in image-availability afforded by digital imaging coupled with delivery via the internet has immeasurably enhanced every aspect of

¹ Roughly 10% of men are fully or partially colourblind. The condition is hereditary and sex-linked. Because the gene is carried by the X-chromosome vastly more men than women exhibit colourblindness. For more information see Bailey and Haddrill.

² Cf Melissa Terras’s chapter in this volume on “Artefacts and Errors: Acknowledging Issues of Representation in the Digital Imaging of Ancient Texts”, 43–61.

³ DIAMM imaging-quality guidelines may be accessed here: <<http://www.diamm.ac.uk/techinfo/quality.html>>.

research that relies on contact with primary sources, and has provided us with surrogates that can be exactly duplicated (as far as we are able to determine) *ad infinitum*, without deterioration or degradation of the original or its copies. Unlike hard copies the digital original cannot become scratched or creased with use; it will not fade or change colour with age (again, at least as far as we are able to determine) and we can carry thousands of them around with us wherever we go in containers smaller than a paperback novel or, in the case of users of DIAMM, by accessing a web URL from any computer connected to the internet.

1.3. Dealing with Damaged Sources

Because the early years of work in DIAMM were primarily concerned with sources that were themselves damaged, one of the earliest exploitations of the images the project acquired was digital restoration. When DIAMM was conceived the idea of digital enhancement existed, but it had never been seriously explored or attempted with manuscript sources. There was no commercially-available software specifically designed for this sort of activity, and there still is not, since several software packages have all the necessary tools available along with all the other extras and gadgets designed for the artistic exploitation of digital image media. The decision was taken at the start of the project to use only existing established commercial software that would not involve the project in software development or maintenance. Adobe Photoshop was at the time the best package, and is still probably the most widely-used image-processing package.

The early fragments digitized by DIAMM had undergone a variety of distressing and damaging experiences that had left them in many cases unreadable, and in some cases barely recognizable as having ever shown music.

Most common in the variety of misfortunes to befall our fragments was their re-use as binding materials, either simply by re-using leaves as wrappers, or more destructively by cutting them up to make quire guards or using them as paste-downs or binding reinforcement. Many of the fragments were caked with glue and darkened by long contact with leather turnovers. Those used as wrappers were often so rubbed that no ink was apparent on the outer face at all. Whole manuscripts survived as distressing palimpsests, with only tantalizing glimpses of musical notation visible at the extreme edges of the over-written material, while others had vanished altogether, discarded during rebinding programs, and surviving only as offsets on the preserved boards to which they had originally been stuck down. That many of these fragments had been found and identified as music at all is a testament to the tenacity of one of the founding directors of the project, Andrew Wathey.

Faced with the publication of the fragments in a long-standing series of b/w facsimiles of medieval music sources, the directors decided that publishing an accurate representation of a nearly-black page was a pointless exercise: any detail that was

apparent in colour would be lost in monochrome. The result was what might be described as a completely new discipline in palaeography: digital restoration. The idea gained currency as fast as digital imaging, but has grown without constraints. Many libraries gave us early pointers to the dangers of digital ‘tampering’ by expressing concerns that a ‘good’ restoration of an image might lead visitors to expect the original leaf to be in such a condition when seen in reality. Encountering the reality might then lay the owners open to accusations of poor husbandry, so in their eyes digital restoration had to be reconsidered as a form of misrepresentation rather than simply as an enhancement to research. This is not a factor particularly considered by those undertaking digital restoration for their own needs, but was certainly an issue for DIAMM when delivering edited images to our users of documents that belonged to other institutions.

1.4. Ethical Enhancement

The first step in enhancement is to decide at what point the process is simple adjustment, and at what point that crosses the line into editing. If an image is correctly captured—correct lighting and exposure; correct colour balance; sharply in focus at the finest detail level—and is therefore the truest reproduction possible of the original then there should be no need for post-processing adjustment to make it look more like the original. If the original source is unsatisfactory but the image of it is ‘perfect’, then *any* adjustment must be deemed to be editing.

With the reservations expressed by libraries in mind, I looked at ways in which to make a reproduction of a ‘restored’ original visibly obvious as an altered source. Captioning is necessary naturally, but a caption cannot express the extent to which the image may have been manipulated or adjusted. Quite simply the only way to demonstrate the extent to which an image has been ‘improved’ is to publish the two versions side-by-side in whatever medium they will appear. Any restoration appearing in print has the added problem of the authority that print confers: a published restoration may be seen as the only, or best, ‘solution’ to reading the document. The examination of these issues led to a re-appraisal of terminology: ‘restoration’ implies the return of something to an earlier state, and we could not say with any reliability that what I did to an image restored an earlier state of the source. The first step therefore was to refer to adjusted images as ‘digitally enhanced’. My first efforts at digital enhancement concentrated on ‘restoring’ faded colours or darkening: so faded ink was darkened and accumulated dirt was lightened, increasing, or in some cases reversing the contrast of the original while retaining its basic colours.

The result was often something that could have been mistaken for the present state of the original document if seen in isolation. In some cases the change was so radical that a user might discard the true image of the original as a mistake of some sort. A better

solution to this problem was to restore in bright colours: instead of selecting faded ink colour and darkening it, I selected the ink colour and replaced it with a completely unlikely colour such as green or purple. The pixels selected were the same, but there was no possibility that the end result could be mistaken for anything apart from an edited image. An unexpected benefit of this was that the contrast of the new colour against the old (green on brown instead of dark brown on pale brown) enhanced clarity and made reading the text much easier. The Stratford example (below) makes use of this technique.⁴

There are difficulties in some cases in deciding which items on a leaf are remnants of text. Enhancement requires this decision early on, and is where the first major mistake can be made that will compromise the results. The first step in restoration of a source like this should therefore always be a process that is applied to the whole image, without any selection of areas or colours. The simplest and least controversial method for this first process is a level adjust, where dark colours are made darker and light colours lighter. For many sources this is extraordinarily effective, and may be the only process required.

In the example given in Fig. 1 the centre gray and white sliders beneath the histogram have been moved to the left, lightening the pale highlights and the mid grays. The black slider has been moved to the right to darken the dark highlights—in this case the writing that we want to see.

Having enhanced the colours that are there, it is usually much easier to determine which belong to text, and these colours can then be selected and enhanced (usually darkened), although once again it is essential that first level enhancements of this type use selections that are applied to the whole image, to ensure that the electronic selection is not predetermined by what the editor hopes or expects to see.

The Stratford wrapper (Fig. 2a-b) is a very good example of how following this practice yielded results that were not those we expected. According to the cataloguing of this item, there is music inside the wrapper, but nothing on the outside. The first image of the back of the outside of the wrapper, when enlarged, revealed what appeared to be the shapes of musical notes in the top left corner, and what appeared to be a chain of ascending notes about a third of the way down in the middle of the page (just below the later writing). A basic level-adjust clarified that the notes on the top left corner were indeed music notation. Using these notes as colour selection reference points, the difference between the colours of the ink and of the background dirt, which were too similar to be separated by eye, was exaggerated, with surprising results. Most interesting was that the line of ascending note-heads in the middle of the page were

⁴ Digital restoration techniques used by the author are described in the DIAMM Digital Restoration Workbook, available for download at <<http://www.diamm.ac.uk/redist/pdf/RestorationWorkbook.pdf>>.

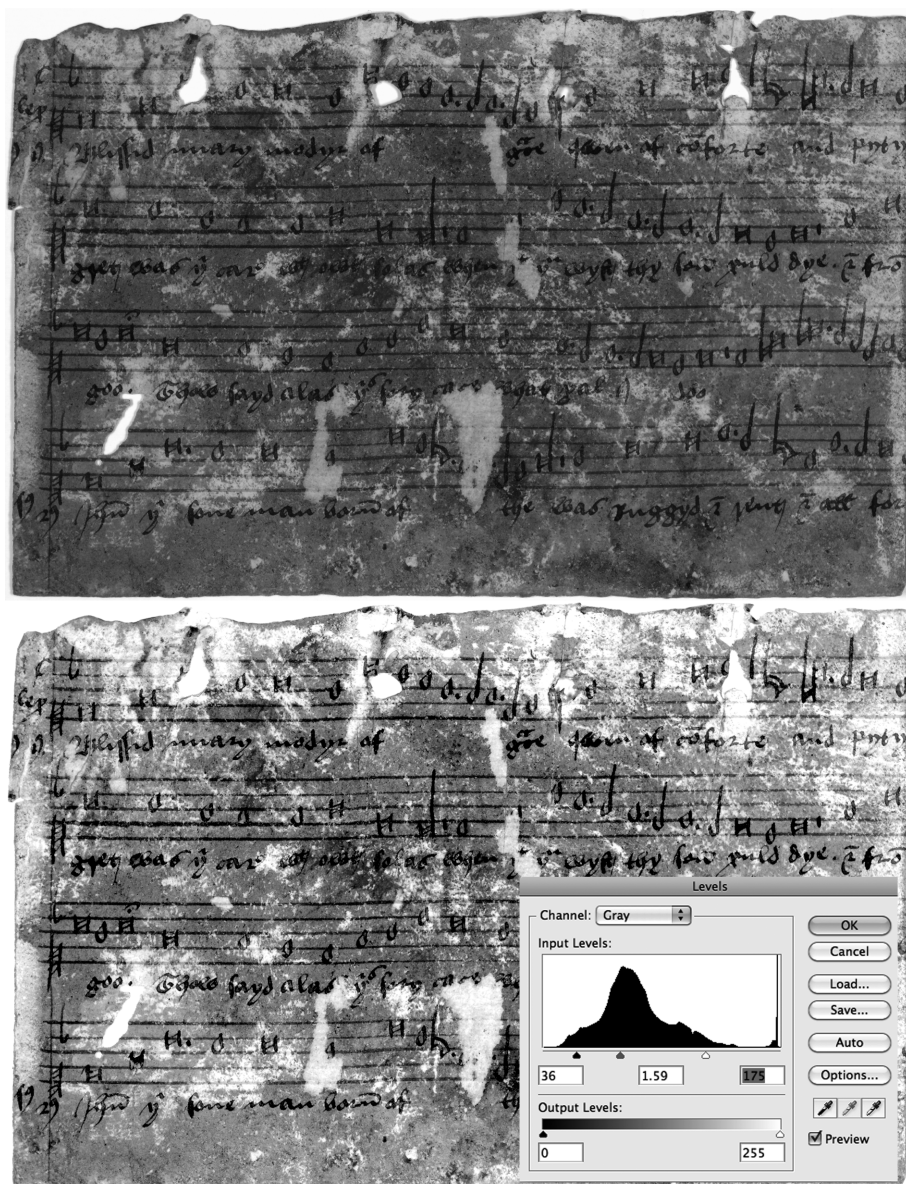


Figure 1. Cambridge, Pembroke College MS 314 supplementary envelope C, fol. 1v.

clearly not part of the text on the page itself, and may have been offset from a facing page, now lost.

Another sample of selection based on what was visible, rather than what we thought might have been there revealed a section of text (a second text to that underlaid beneath the music) that had gone unremarked even during the early stages of the restoration process (see Fig. 3).

The final example (Fig. 4) is of a leaf that was also used as a wrapper, but was not quite as damaged as the first sample. Various stages in the restoration process are shown, but it is difficult to decide at what point this leaf has been over-restored. Whole-image processes on the leaf revealed significant material, but although further processes improved some areas of the document, it disimproved others. The end user therefore needed a fluid version where some layers of processing could be turned on and off (easy in Photoshop, where the restoration is always done using 'layers'). In order to present a finished result that could be printed or shown outside Photoshop processes had to be applied only to selected areas of the document. The result is not a global adjustment process, but one that has had selective judgements applied, and thus is more of an editing process than a restoration.

1.5. Unethical Enhancement

I recently had cause to re-examine the ethics of digital enhancement when I was asked to photograph a document on which the owner was convinced certain words were written. I could not see these words, but the owner was determined they were there, and wanted to have the right sort of pictures, and to be taught how to digitally enhance them, so that he could prove what he passionately believed. The correct way to handle this situation would be to hand the picture to a third party who knew nothing about the content and ask them to see if they could find anything on it. The chances are though, that the third party would find nothing because they had no idea what they were looking for. At some point there has to be a decision about what type of enhancement to do, and that depends on what you may find. It helps if you know there is 'writing' there, but you don't know what it says: that way the writing can be made more visible by the disinterested party, and then it is up to the expert to read what it says and present the image to others for their opinion. This is not very different from the old-fashioned magnifying glass and palaeographer's trained eye: ultimately we have to trust the expert.

Enhancement that is undertaken to make an image show what cannot be revealed by 'justifiable' methods is no longer enhancement but editing. A colleague dubbed the results of this type of work a 'fake-simile'. Editing (or unethical process) is far from undesirable. Sometimes the results are sufficiently illuminating that the process is entirely justifiable, though how far to go in the pursuit of this process can present

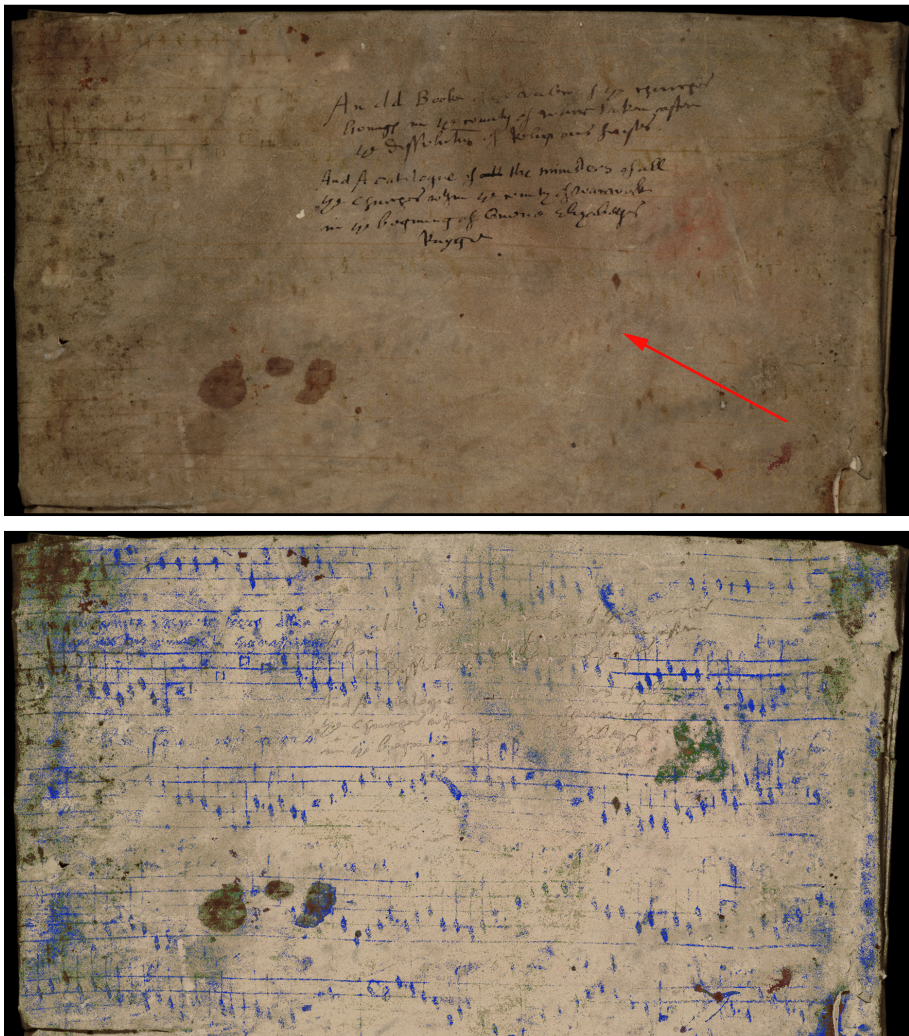


Figure 2. GB, Stratford-upon-Avon, Shakespeare Birthplace Trust, DR 37 Vol. 41, part of the back cover before and after restoration (suspect notes indicated by arrow).



Figure 3. GB, London, British Library Add. Ms. 41340 fol. 1 before and after restoration, © The British Library Board.

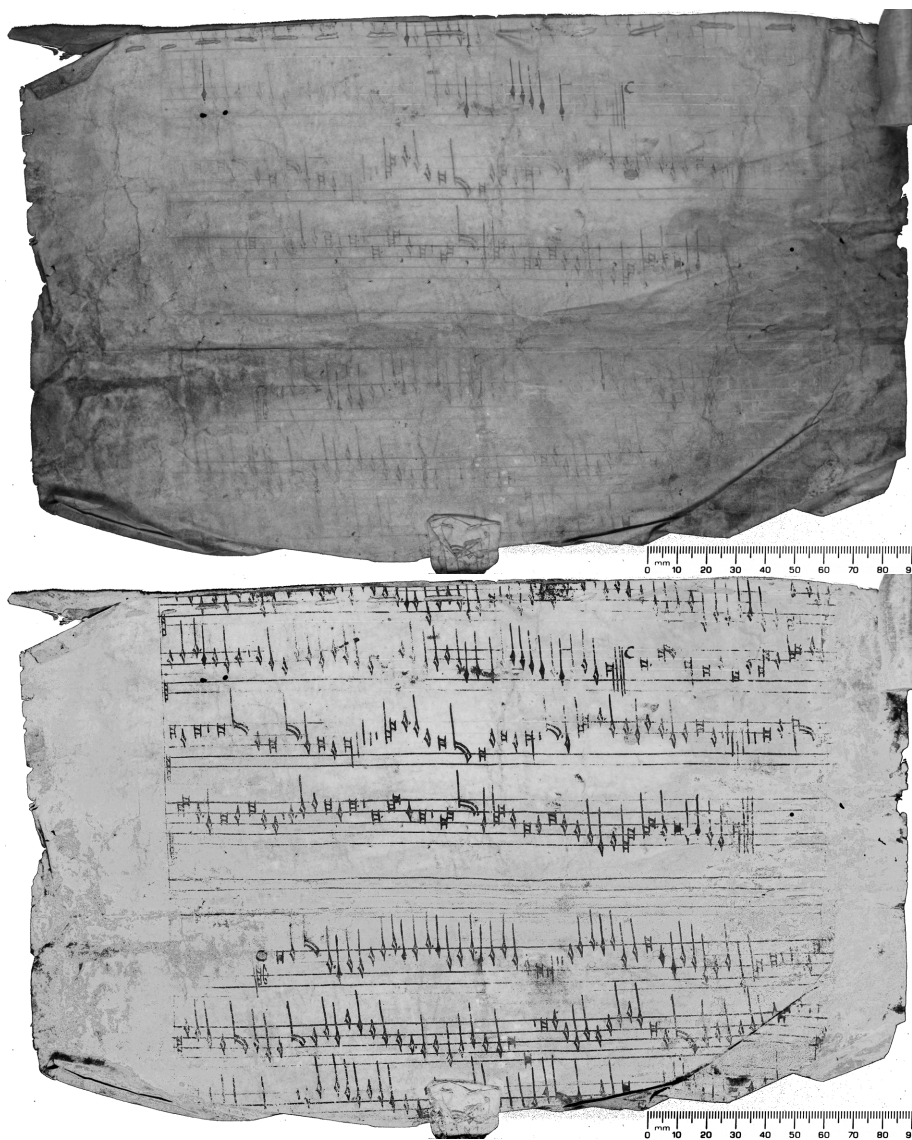


Figure 4. USA, Princeton MS 138.41 (Preece-Leverett fragment), photographed while in the private collection of Christopher de Hamel.

another stumbling block in presenting those results to a wider readership. As long as the restored version is always presented alongside as near a perfect representation of the original as possible, then the reader is still in a position to judge the veracity of the restoration.

The following example was taken from a paper manuscript damaged by ink burn-through. The first figure (Fig. 5) shows the original state of the manuscript. By magnifying the image on screen it was possible to see the difference between notes written on the reverse and those on the front face. Both notes have ‘fuzzy’ edges, but those on the front face have a more solid dark core. The duct of the script also gives clues to which notes belong on which face of the leaf. We can remove the notes that we are certain do not belong, and doing that removes considerable confusion from the visual field, making it possible to decide on a further group of notes that can be ‘removed’. The process is repeated, also allowing for grammatical probability—so that where there is a choice between a next-door note and one at some pitch distance from the ones that we are sure of, the most sensible choice is the next-door note. Where there is heavy blurring, the position of the notehead can often be determined by using the stem, which—despite being hand-written—is usually of a uniform length. This process is best undertaken by someone who at least understands the vocabulary of the period, since that allows the restorer to determine which shapes are notes or rests, and which are other grammatical marks or nothing to do with the music. The reverse face can be restored simultaneously so that it can also be used as a reference for the repair of its partner.

At this late stage in the process working with the ‘expert’ musicologist in tandem with the restorer is necessary, since the expert knows the grammar of the period and the syntax of the musical work, and can make more informed decisions about which notes to remove and which to leave. However allowing the expert musicologist to do the work themselves from the start has proved inadvisable, since the temptation to see what they want to be there is often very hard to resist!

The second figure (Fig. 6) shows the ‘repaired’ version of the page. All but three notes on this page were restored with sufficient certainty that they were not in question. The three remaining notes were more difficult: by reconstructing the whole piece of music it was fairly clear what those obscured notes must be, but the ink that was apparent on the page did not match that expectation. At this point a heavy editorial hand was applied.

The process involved here is very different from that of selecting a colour and having the software find all instances of the same colour on the leaf. Since both foreground and show-through inks have the same values this technique of colour separation does not work. The process employed here is cloning—akin to simply rubbing out or deleting the colours that are not desired—where a clean part of the work of the same scribe with matching lines and line-spacing is copied over the area that we want to obscure. A leaf

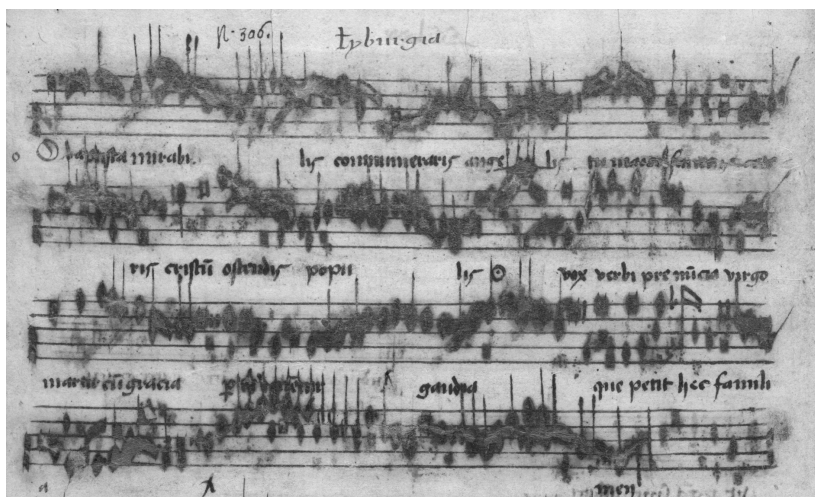


Figure 5. Italy, Museo internazionale e biblioteca della musica di Bologna, MS Q. 15 fol. 309v lines 1–4 as it appears in the MS. (reproduced by kind permission).

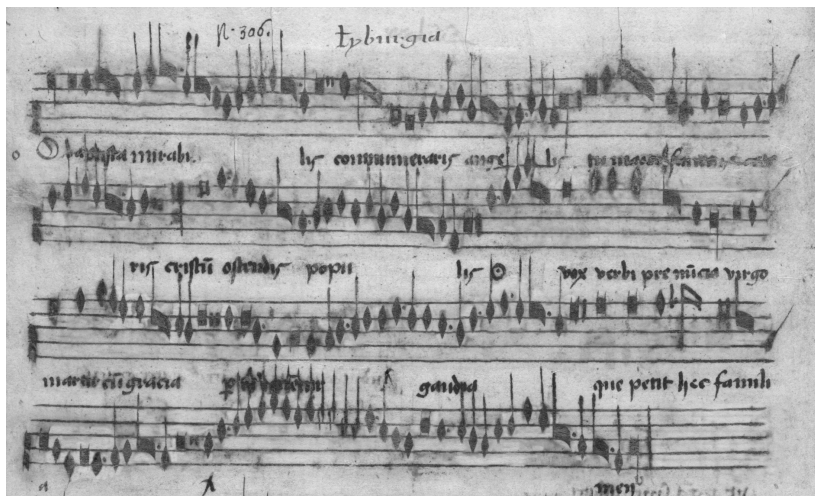


Figure 6. Italy, Museo internazionale e biblioteca della musica di Bologna, MS Q. 15 fol. 309v lines 1–4. Restored 'fake-simile' version, digitally altered by cloning from cleaner pages written by the same scribe.

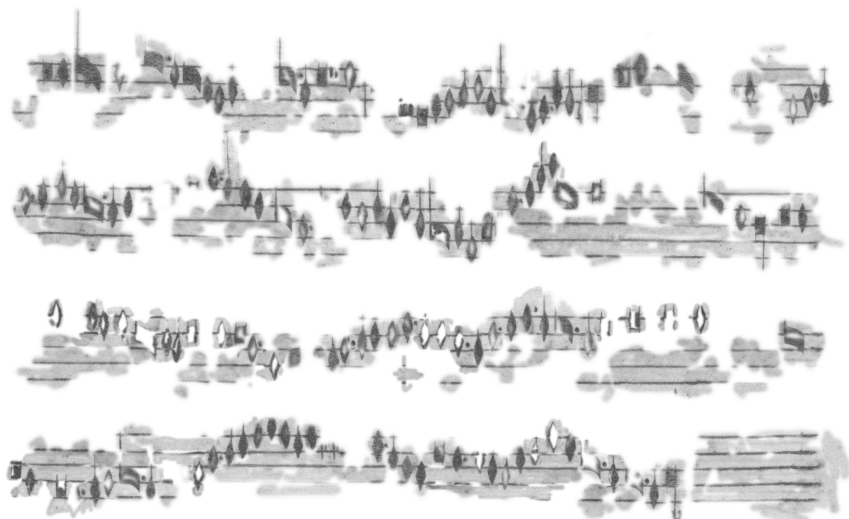


Figure 7. MS Q.15 fol. 309v lines 1–4, cloned details only with base image removed.

from work by the same scribe that was undamaged (and cleaner areas of this page) was used to supply undamaged versions of notes, and clean sections of stave.

The final figure (Fig. 7) in this set shows only the cloned areas of the leaf, with the main leaf image removed. This shows what was overlaid on the main image to create the cleaned-up result. It was not particularly meticulously done since the object was readability, not perfection: a more elaborate restoration would eliminate the obvious points of changes in paper colour, and would clean up the remaining blurred notes that have been left here since work was not required simply to make them readable.

Cloning though is such a dubious activity that it cannot be called restoration (unless it is called ‘unethical restoration’), since does not adjust information that is already there, it replaces it. In a moment of idle devilment while writing this article I inserted the word ‘elephantēs’ into an unsuspecting line of biblical text from a medieval bible leaf (Fig. 8). It was not particularly carefully done and took about 5 minutes. It is included here only to demonstrate how easy it is to mislead ourselves and others when tampering with digital images.

The final example (Fig. 9) shows a process that is not so much unethical as simply not restoration, just the use of image-processing to elucidate what the eye and the skills of the expert musicologist can determine. The subject was a scan of a conventional b/w

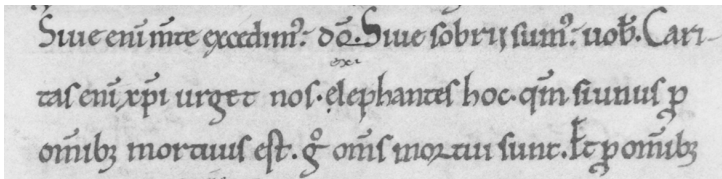


Figure 8. Digital fallacy: cloned text on a medieval bible leaf.

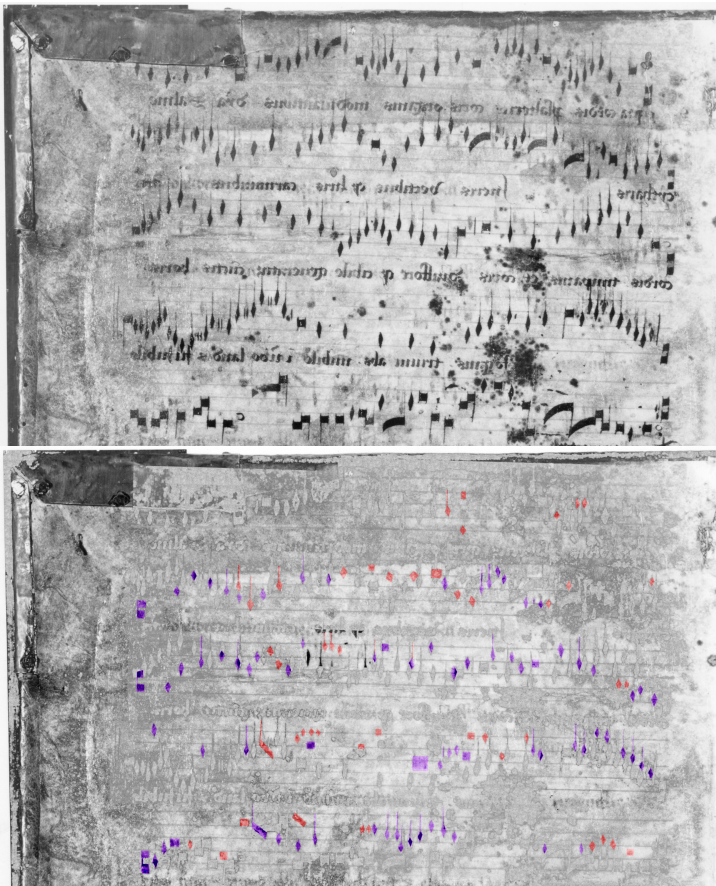


Figure 9. USA, Cambridge, Harvard University, Houghton Library, fms Typ 122, fol. A, original and restored versions. (The first image has already been flipped.)

photo that was taken when a paste-down (still stuck onto a binding board) was treated with a chemical to make the parchment more transparent thus making the reverse of the page visible. In this case we want to ignore the notes on the superior surface, and enhance the ones on the reverse. The document is 'flipped' giving a mirror image, so that the notes we want to see are facing forwards, and the confusing notes that we want to ignore are reversed. Here again a trained eye is useful: the restorer in this case first 'removed' all the notes that he could see by duct and shape were on the front surface. He then flipped the document so that the reverse-face text was correctly orientated and enlarged the document on screen. Where he found shapes that he recognized as musical notation that he was sure did *not* originate on the front face he selected and then 'colorized' them purple. Notes that he was fairly sure belonged on the reverse (or that he considered were of dubious value) were coloured red, and it is evident from the second selection that there is some conflict between the two readings. The result was sufficient to identify the piece of music.

1.6. Delivering Metadata

The delivery of 'improved' images alongside damaged originals in the DIAMM content is a useful adjunct to the simple study of high-quality digital reproductions. However, key to examining and re-examining sources is easy access to existing secondary materials and metadata. As well as providing searchable electronic versions of the two seminal catalogues in medieval and early-modern musicology, DIAMM also provides page-images of the catalogues, allowing users to page through the books in their original typeset form. The cost and extent of the catalogues has, up until now, required those wishing to consult them to visit a research library:

CCM *Census Catalogue of Manuscript Sources of Polyphonic Music 1400–1550*
(5 vols.);

RISM *Répertoire International des Sources Musicales: Manuscripts of Polyphonic Music c.1000 to 1550, Series B IV* (5 vols.).

There have been minor supplements to RISM Series IV, but essentially these two books are the only reference works that attempt to give a consistent level of information for every known manuscript of Medieval and Early Modern music in the world. 'Attempt' may be the right word, particularly in CCM: because the entries were compiled by local scholars, their descriptions of the documents tend to reflect their research interests, so one manuscript will have a long provenance history and no palaeographical information at all, while another will have extensive codicological information but no discussion of the musical contents or provenance. RISM is slightly more consistent, and provides invaluable inventories for all the manuscripts which the editors of CCM decided not to include (probably for reasons of cost).

At present the delivery of metadata through the DIAMM website is very limited but the full content of the back-end working database will be brought online during 2011 (possibly by the time this text is published). This will allow users to search on any or all of the following by using directed browsing as well as search forms:

- The full text source-description field;
- Text incipits;
- Full text transcriptions for many sources (in original and standardized spellings);
- Language database;
- Genre;
- Work title in original and standardized spelling;
- Voice part designations;
- Clefs and clef combinations;
- Scoring;
- Composer.

Inventories of each manuscript link to the images of each item together with editions of the works where those exist, and bibliographical links to articles relating to individual pieces. The manuscript metadata already links to bibliographical items (as does metadata for individual works within each source), but will also now link to a ‘person’ database giving scribes, dedicatee, binders, authors (of treatises), decoration style models, establishment patron, transcriber, owner etc.; manuscripts are also linked together in sets such as those manifest in the structure of the books (e.g. individual books in partbook sets, fragments of dismembered books) and other sets of a more intellectual type: links by copyist, illuminator, scriptorium, patron etc. will facilitate searches by or within groupings of manuscripts.

Musicology projects are increasingly working together so that work and resources used to create one database or delivery system are both available to other datasets and also transferable between datasets. There is a long-standing policy of sharing and collaboration between a number of large database-driven projects, and this will eventually become a considerable collaborative international resource in a way that would have been unthinkable only a few years ago. The participants: DIAMM, CMME, The Motet Database (which has been incorporated into the DIAMM database), Base Chanson (Ricerca) and Die Musik des Trecento (DMT), all deliver their data free to any user. Naturally cross-searchability of these datasets would be an enormous step forward, and is the goal of all the projects involved in a pilot phase of work to create a services model known as REMEDIUM.

The Motet database, Base Chanson and DMT are ‘classic’ metadata-driven resources, but CMME—Computerized Mensural Music Editing—had different origins for its dataset.

2. CMME

Theodor Dumitrescu began his work on creating an editor for early music notations while a Computer Science student at Princeton University, and continued to develop it later with research fellowships in Tours and Utrecht. He is currently a lecturer in the Music Faculty at the University of Utrecht, where development of more completed and polished versions of the software continues. The project is now able to deliver an accessible and intuitive version to researchers, as well as customize the software for specific projects (Dumitrescu; Selfridge-Field; Veit).

Mainstream music processing software is expensive, extremely complex and ultimately unsuitable for diplomatic transcription of medieval and early modern notations. Even if transcribing into modern notation this software is often too inflexible to allow the use of specialist editorial policies, and is also primarily designed for print output, not for online delivery. More important, these softwares are not XML-based, so input content is therefore not searchable. The online environment is particularly appropriate for the fluid presentation of materials and information when a single fixed presentation is undesirable. This is particularly true in a complex variorum environment that deals with more than a few primary sources although even comparing small numbers of sources print cannot be used to highlight different aspects of the variation between sources chosen by the user, only those that the editor has predetermined. Increasingly, and particularly with early music repertoires, the needs of musicology require searchable content, and for the first time in music-processing this is now deliverable. Dumitrescu describes the purpose of his software thus:

Without the need for a single fixed visual form on a printed page, a computerized edition system can utilize one editor's transcription to create any number of visual forms and variant versions. The result is an entirely new form of critical music edition in which dynamically generated, user-configured formats remove the unwieldiness of multiple printed editions, replacing it with the concept of multiple states of a single edition. The early music editor's task returns to the truly critical aspects of interpreting the text, rather than the ultimately unsatisfactory process of making presentation decisions which must limit the usability of the edition.

(Why online editions?—CMME website)

CMME, as well as offering the facility to see music transcribed in a number of different editorial modes (chosen by the user/reader), can also be used to demonstrate variants between sources that can be chosen by the user, and highlighted with different visualisation techniques both on the score and in a configurable critical apparatus window (Dumitrescu and van Berchum).

As well as providing a clear WYSIWYG (What-You-See-Is-What-You-Get) interface for point-and-click editing using the same notation as in the original source (including red notation), a text editor, commentary tool, and support for notation features such as arbitrary proportions and colors, the viewer for finished editions includes an applet that allows the viewer to choose how they wish their score to appear. It is possible to view the music in separate parts as it appears in the original source, or as a score.

Having designed the software, Dumitrescu and his team are now utilizing it to create a set of dynamic electronic editions of important printed and manuscript sources. As well as serving as test materials for the software, this corpus represents a major contribution to the scholarly materials available in musicology:

The music editions which populate the CMME corpus do not represent an anonymous mass of information punched in by disinterested data-entry workers; these are fresh editions, produced from primary source materials by musicological experts and kept at a high standard by our international board of editorial advisors. The reader of Shakespeare or Chaucer expects an edition which has been prepared with great care and knowledge acquired through study and experience, as does the performer of Lassus motets. There is no reason to waive these same requirements in online editorial endeavors.

These principles of scholarly arbitration and peer review remain an important element in the formation of CMME “Editorial Projects.” Corresponding in many ways to a volume in a printed edition series, an editorial project gathers up a collection of related compositions to be presented under the guidance of one editor or editorial group—for example, the contents of a single manuscript, the complete works of one composer, or a set of pieces known to come from a single court or city. As with any publication of early materials built on sound scholarly standards, a CMME edition provides the user with an introduction by the music’s editors illuminating the historical, musical, and analytical context of the edition’s contents. The web pages for individual compositions offer further commentary specific to each work.

Beyond the set of editorial projects, however, users will quickly encounter a wider array of information in the Database section concerning musical sources, composers, and compositions which are not yet represented by music editions in the CMME. This network of contextual information [...] allows readers to explore the broader environment of the music editions in the corpus. The CMME meta-data collection limits itself largely to the one major element which is missing from the standard reference works on 15th- and 16th-century sources: the listing of actual contents, giving the names and composers of all compositions and their locations in the sources.

(Edition projects and “meta-data”—CMME website)

Databases, however useful, are in a way an unexciting adjunct to codicology and palaeography: the concentration in current research projects on creating and populating them tends both to force a traditional (and necessarily) hands-on discipline to become distant and computerized, and further to imply that we no longer need to visit the original document, a fact that is manifestly untrue. Applying for funding simply to spend time with a source is nowadays almost unthinkable when building a case for a major research initiative. Both peer- and funding-pressure push us to create digital resources that we—and the technology available—may not be ready to create. We have only to look at the rapidity with which software has been created to ‘read’ historic newspapers and make a searchable archive from them to realize that almost anything we do now, with painstaking effort, may be automated in five years time. A very good example of this is Olive Software’s *ActivePaper* which reads and classifies newspaper content, producing a very accurate searchable result.⁵

Many projects, but more significantly many funders, are seduced by the digital world, and the skills of manuscript handling, study and description are in danger. For many new students, if the information is not on the web then it doesn’t exist, and that means that resources like DIAMM, for all the good they do, are also skewing study towards those documents that are accessible online, and away from those that are still difficult to access.

3. GAMERA

A case of research being driven by technology is to be found in the software creation of Ichiro Fujinaga, Assistant Professor in the department of Music Technology at McGill University, Montreal: Gamera. Gamera is not a packaged document recognition system, but a toolkit for building document image recognition systems. It makes the development of a new recognition system quite easy, though this still requires some time commitment. Gamera is a cross platform library for the Python programming language. Apart from providing a set of commonly needed functionality for document image analysis, Gamera additionally allows for custom extensions as C++ or Python Plugins and as Toolkits.⁶

I have often been surprised by listening to colleagues (particularly in classical and medieval subjects) wishing for a software that would recognize letter shapes in manuscript hands, when this is precisely what Gamera will do, and it has been around for some time and has been quite widely exploited. The software is teachable,

⁵ An example of the software in action can be found at the online archive of *The Scotsman* newspaper.

⁶ For a lengthy list of publications describing either the Gamera project itself or other research projects performed with the use of Gamera, see the list of publications on the Gamera website. Of particular interest to non-musicologists may be: Reddy and Crane; Canfield; Droettboom; Droettboom, MacMillan and Fujinaga.

so that when it finds characters it does not recognize it will suggest what it thinks they are, then when confirmed or corrected by the user it will re-evaluate the content of the document under consideration based on its revised database. Important here is its ability to *suggest* a reading, since sometimes a human reader will see a letterform and assume what it is from context. While this is often an effective way of dealing with damaged forms, or those obscured by interference of various types⁷ it can mislead the reader into a completely incorrect interpretation of the material (Bowman, Tomlin and Worp).

The early development of Gamera arose from a desire to create a software that would read printed lute tablature. Tablatures are one of the most ancient forms of musical notation, showing players where to put their fingers, rather than the notes that will sound (which is what modern notation does). Lutenists still play from tablature, and some other tablatures are also occasionally still used.

Printed lute tablature has various features that make it both difficult and easy to deal with electronically: it is quite regular (the print was made from wooden type); the musical ‘events’ are evenly spaced, and there are a relatively small number of symbols in comparison with normal musical notation. In ‘French’ tablature six lines are used to represent the strings or courses on the lute. The top line represents the highest-sounding course. Letters are placed on the line to indicate which frets are to be stopped: the letter ‘a’ is the open string, ‘b’ the first fret, ‘c’ the second fret and so on.⁸ Rhythm is indicated by a series of flags above the staff and, unusually for music in this period, the music is divided up into regular periods by barlines. The difficulty in teaching tablature to a machine is that because the letters printed on the lines are made with single-impression movable type, there are sometimes—but not always—gaps between one block and the next. The first task therefore was to remove the horizontal lines (deskewing the document along the way).

As hand-carved wooden type is prone to both natural irregularity and damage from use, the symbols are not very regular. In the first screenshot of the classifier below (Fig. 11; normally this appears in colour) the shapes in pale gray have been defined according to the information in the Gamera database. The shapes on a dark gray ground are those which the software is questioning. If the user clicks on a symbol its position is shown and highlighted in the document (in the bottom half of the window) so that it can be checked by context if it is not immediately obvious what it should be. Note that

⁷ E.g. when attempting to read text on the remains of wax tablets, where the wax has gone and all that is left is the indentation of sometimes several layers of writing on the tablet base, where interference is caused not only by the layers of writing, but also by the grain of the wood; see Vindolana Tablets Online.

⁸ Italian tablature reverses the lines, so that the bottom line on the staff represents the highest-sounding course (thus matching the lines to the position of the courses), and the frets stopped are indicated with numbers (0–9) rather than letters. Obviously this system ran into trouble if courses higher than the 9th were required, and it extended into letters (x, etc.) from that point. Fortunately very high frets were rarely used in music of the period of the lute’s heyday.

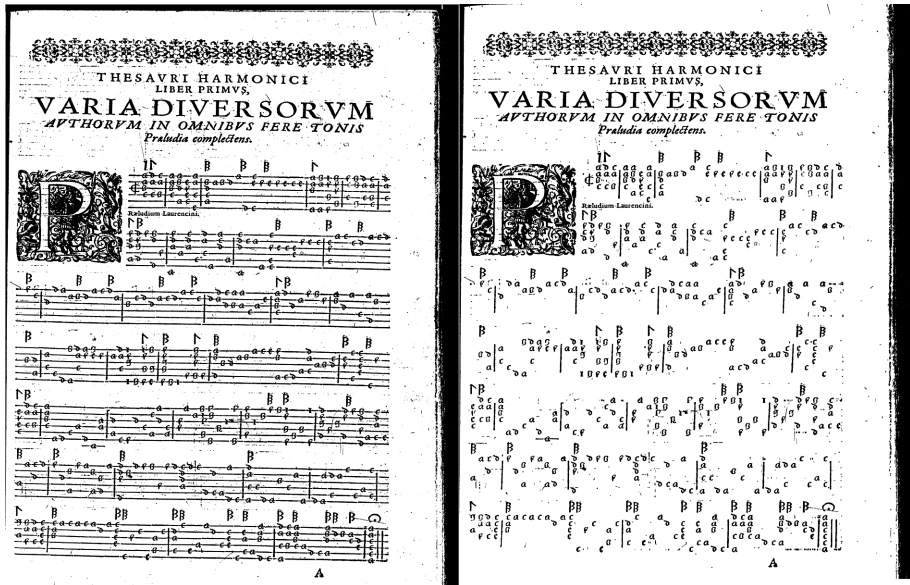


Figure 10. A page of printed lute tablature before and after staff-line removal.

although the symbol is being questioned, the software has nevertheless placed it in the correct category.

The user, having decided what the symbol is, assigns it to the classification list in the left part of the window, and the recognition database is updated accordingly.

The same process of line-removal can be applied to other musical notation such as neumes, commonly used to notate chant (Fig. 12).

The classifier database for that document set behaves in exactly the same way as for the tablature. The neume version is a significant step forward from the printed tablature, since this text is handwritten, and far more irregular than the tablature, since the scribe is compressing some figures to fit within a certain space. The spacing here is very irregular, but that does not cause a problem for the software. The neume shapes however are equally irregular—the first and second rows in the classifier show at least two different forms of each neume shape, but Gamera has successfully classified them, and only needs to have one confirmed.

The final example, from a handwritten Greek text was used to show that any shape system can be recognized, as long as the database has been given the basic set of symbols.

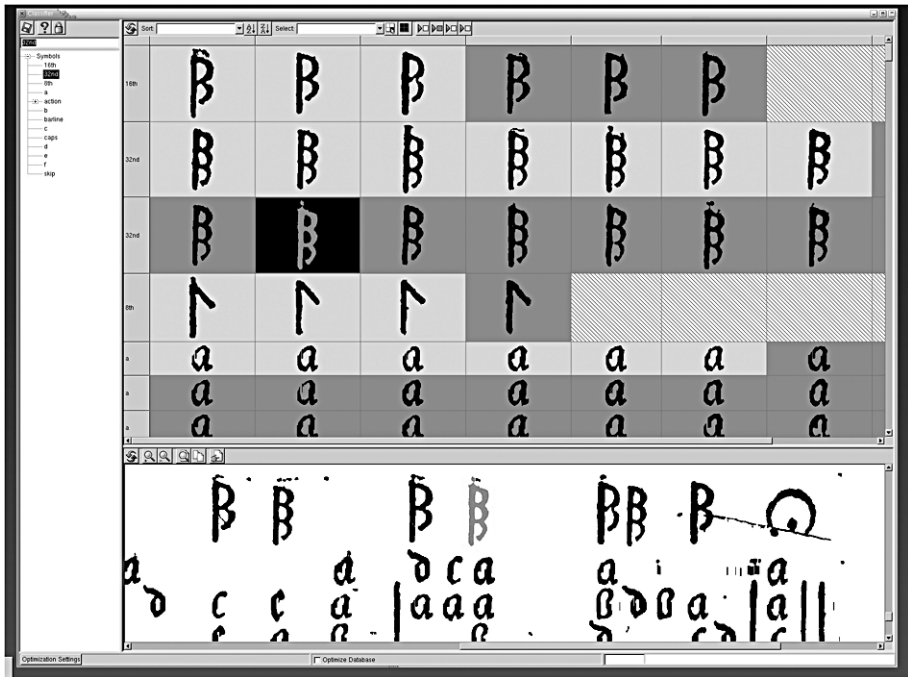


Figure 11. The Gamera classifier window, showing the identification of symbols on the pages shown in Fig. 10.

A custom version of Gamera was employed in the Online Chopin Variorum Edition (OCVE) where a massive number of digitized printed pages had to be marked up into individual bars. Although printed, there is very little regularity in the layout: in order to optimize page-space the engraver would compress the space between stave lines, and bars are never the same width. Using the horizontal line-removal feature, and then by creating a custom recognition system for vertical lines Gamera was used to automate markup of over 7000 pages of music, creating a list of co-ordinates for the bar lines that was then used to generate individual crops of each bar on each page, allowing a user to select one bar (or group of bars) and see the same bar (or set of bars) from the same piece from all the editions available in the dataset.

Fujinaga also prepared a custom version of Gamera for DIAMM, in which he input several hundred JPEG images of manuscripts in our collection, and programmed the software to examine the manuscripts and identify scribal concordances or instances

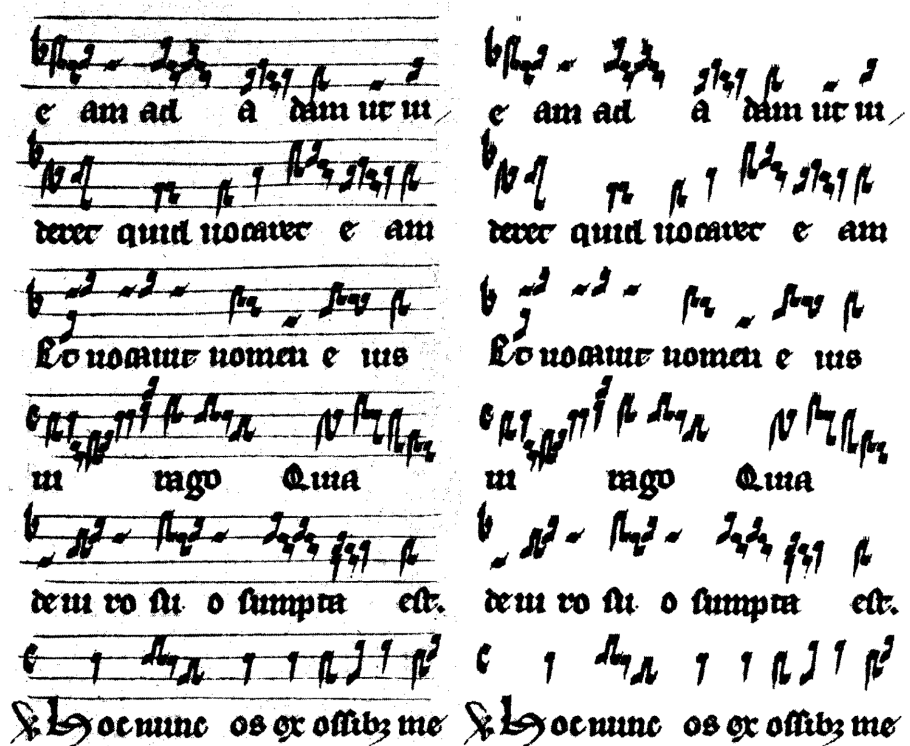


Figure 12. A page of handwritten neumes before and after staff-line removal.

where the scribe was similar. The software delivered a set of results that ranked the manuscripts by their similarity to the primary source expressed as a percentage (Fujinaga). Since it had scanned *every* source in our repertoire it had dealt with a number of materials that would have been extremely difficult for a human to handle, and the results it delivered were extremely convincing: it identified scribal concordances that we already knew about, and correctly ranked materials that were similar so that the expert palaeographer was directed to all the relevant documents in the collection without having to eyeball the entire collection themselves. Further refinement in the image collection to ensure accurate size reproduction for each sample would have enhanced the quality of the output further. Clearly the software has tremendous potential for exploitation in the fields both of script and shape recognition, and the types of study that

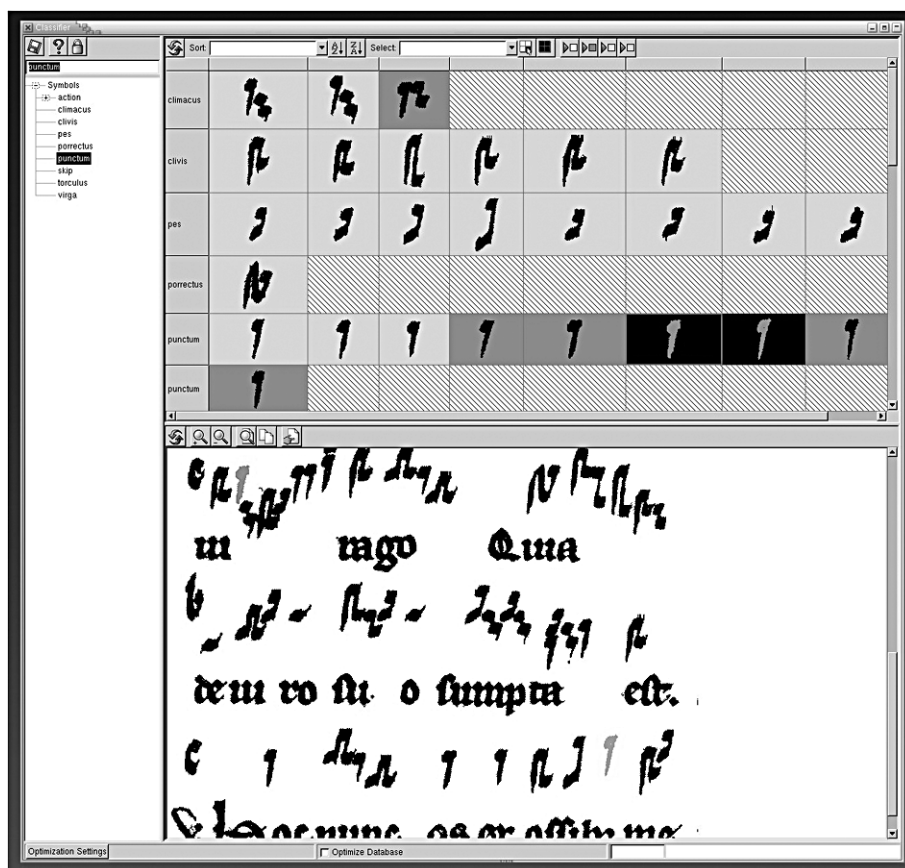


Figure 13. Gamera classifier window showing the identified shapes on the page in figure 12.



Figure 14. Gamera classifier window dealing with a page of Greek text.

can only be done by the analysis of huge bodies of data, most efficiently and accurately managed by computer.

4. Changes to the Discipline

Pre-digital palaeography has relied on the memory and the reportorial breadth of individuals: the visual memory of the distinctive characteristics of a scribal hand—and the memory of where that hand appears—is necessary to the identification of scribal concordances; in musicology an aural memory of entire musical works across a massive musical spectrum might be necessary to identify a newly-discovered fragment. Although a search can be narrowed down by using physical features such as notational idiosyncrasies, this can be misleading if the source was copied outside its original region

or date period; in every discipline everything we read or see has to be retained if we are to make the best use of both primary and secondary, and we have to read or see everything available.

The limits not only of memory therefore, but of the amount of time we have to feed content into it, have shaped our research and study until now. There is a danger that reliance on digital and online content will cause researchers—particularly those new to the field—to limit their knowledge to what is available online, but that danger aside, we are now given the opportunity to search for the information we need across complete data sets instead of limited ones, and to search information that we might previously not have been able to find. In theory, a search for music manuscripts that might have been connected with the Electorate of Saxony can now be made reliably, with every reference delivered, without having to read the entire content of RISM or CCM and possibly miss a few references.

However, the fact that this is possible is dependent on the creation of the dataset and making it available for searching: research is still limited by the availability of content, and inputting that costs time and money, and funding is rarely available simply to create content. Dumitrescu's database of musical sources is searchable, but only the content that he has so far input can be searched. DIAMM will be eminently searchable, but the content is not yet complete, and data input still requires human analysis of, and decisions about, existing data. The creation of new information in the form of catalogue descriptions is even more difficult, since it requires a significant level of expertise, and most of those with that expertise do not have the time to devote to re-cataloguing work.

Fujinaga's software can be programmed to deliver a wide range of results as long as it is fed sufficient information, and its results carefully corrected to create a well-informed shape-database.

These are not the only sources of data available to aid research, but they give an idea of the type of resource that is now becoming, if not commonplace, at least an expected part of modern research. What remains is to populate these mechanisms with data, and that activity is fast moving from time-consuming and expensive manual input to automated systems. We have found that if we wait long enough, someone will find a way to automate what we have to do manually at the moment. The availability—and more significantly—searchability of data is inevitably going to change research, but although we can speculate now, only time will tell in precisely what manner.

Bibliography

- Bailey, Gretchyn and Marilyn Haddrill. *Color Blindness*. San Diego, CA: All About Vision, 2009. <<http://www.allaboutvision.com/conditions/colordeficiency.htm>>.
- Bowman, Alan K., Roger S. O. Tomlin and Klaas A. Worp. "Emptio Bovis Frisica: the 'Frisian Ox Sale' Reconsidered." *Journal of Roman Studies* 99 (2009): 156–174.

- Canfield, Kip. "A Pilot Study for a Navajo Textbase." *Proceedings of The 17th International Conference on Humanities Computing and Digital Scholarship* (ACH/ALLC), 2005. 28–30. Online: <http://gamera.informatik.hsnr.de/publications/canfield_navajo_05.pdf>.
- CCM: *Census Catalogue of Manuscript Sources of Polyphonic Music 1400–1550* (5 vols.). American Institute of Musicology. Suttgart: Hänssler-Verlag, 1979.
- CMME: *Computerized Mensural Music Editing*. <<http://www.cmme.org/>>.
- Craig McFeely, Julia. "Digital Image Archive of Medieval Music: The evolution of a digital resource." *Digital Medievalist* 3 (2008). <<http://www.digitalmedievalist.org/journal/3/mcfeely/>>.
- DIAMM: *Digital Image Archive of Medieval Music*. Oxford: University of Oxford, 2008. <<http://www.diamm.ac.uk/>>.
- DMT: *Die Musik des Trecento*. Hamburg: Musikwissenschaftliches Institut der Universität Hamburg, 2001–2007. <<http://www.trecento.uni-hamburg.de/>>.
- Droettboom, Michael. "Correcting broken characters in the recognition of historical printed documents." *Joint Conference on Digital Libraries* (JCDL'03) 2003. 364–366. Online: <http://gamera.informatik.hsnr.de/publications/droettboom_broken_03.pdf>.
- Droettboom, Michael, Karl MacMillan and Ichiro Fujinaga. "The Gamera framework for building custom recognition systems." *Symposium on Document Image Understanding Technologies* 2003. 275–286. Online: <http://gamera.informatik.hsnr.de/publications/droettboom_gamera_03.pdf>.
- Dumitrescu, Theodor. "Corpus Mensurabilis Musicae Electronicum: Toward a Flexible Electronic Representation of Music in Mensural Notation." *The Virtual Score: Representation, Retrieval, Restoration*, ed. Walter B. Hewlett and Eleanor Selfridge-Field, *Computing in Musicology* 12. Cambridge, MA: MIT Press, 2001, pp. 3–18.
- Dumitrescu, Theodor and Marnix van Berchum. "The CMME Occo Codex Edition: Variants and Versions in Encoding and Interface." *Digitale Edition zwischen Experiment und Standardisierung: Musik—Text—Codierung*. Ed. Peter Stadler and Joachim Veit. Beihefte zu Editio. Tübingen: Niemeyer, 2009. 113–130.
- Fujinaga, Ichiro. "Creation of a Searchable Database for note shapes." *DIAMM. Technical Content and Web Delivery Workshop. 29–30 July 2004. St Hilda's College, Oxford*. <<http://www.diamm.ac.uk/redist/pdf/July.pdf>>. 21–25.
- Gamera Project. Krefeld and Mönchengladbach: Niederrhein University of Applied Sciences, 2008–2010. <<http://gamera.informatik.hsnr.de/>>.
- OCVE: *Online Chopin Variorum Edition*. London: King's College London, 2010. <<http://www.ocve.org.uk/>>.
- Olive ActivePaper Archive. Aurora, CO: Olive Software, Inc., 2010. <<http://www.olivesoftware.com/>>.
- Reddy, Sravana, and Gregory Crane. "A Document Recognition System for Early Modern Latin." *Chicago Colloquium on Digital Humanities and Computer Science: What Do You Do With A Million Books* (DHCS 2006). Chicago, IL: University of Chicago, 2006. Online: <<http://dhcs2006.uchicago.edu/abstracts/reddy.pdf>>.

Ricercar. Programme de recherche en musicologie. [Base chanson.] Tours: Centre d'Études Supérieures de la Renaissance de Tours (UMR 6576 du CNRS).

<<http://ricercar.cesr.univ-tours.fr/>>.

RISM: *Répertoire International des Sources Musicales; Manuscripts of Polyphonic Music c.1000 to 1550.* Series B IV (5 vols.). Munich: Henle-Verlag, 1965 ff.

The Scotsman. Digital Archive. Edinburgh: The Scotsman Publications Ltd., London, 2010.

Selfridge-Field, Eleanor. "XML Applications in Music Scholarship." *Music Analysis East and West.* Ed. Walter B. Hewlett and Eleanor Selfridge-Field. *Computing in Musicology* 14. Cambridge, MA: MIT Press, 2006. 21–40.

Veit, Joachim. "Musikwissenschaft und Computerphilologie—eine schwierige Liaison?" *Jahrbuch für Computerphilologie* 7 (2005): 67–92. Online:

<<http://computerphilologie.tu-darmstadt.de/jg05/veit.html>>.

Vindolana Tablets Online. Oxford: Oxford University 2003. <<http://vindolanda.csad.ox.ac.uk/>>.

Transkription und Textkodierung



Transcription and Text Encoding

Ad fontes – mit E-Learning zu ersten Editionserfahrungen

Isabelle Schürch, Martin Rüesch

Zusammenfassung

Vor über sieben Jahren ging das E-Learning-Programm *Ad fontes* online. Sein Grundziel ist es, Studierende aus den Disziplinen der historischen Kulturwissenschaften mit den Herausforderungen der Archivarbeit vertraut zu machen und ihnen eine Möglichkeit an die Hand zu geben, das Lesen alter Schriften selbständig zu erlernen und zu trainieren. Als eines der führenden webbasierten Angebote in diesem Bereich wird *Ad fontes* heute an mehreren Hochschulen in Ergänzung zum traditionellen Präsenzunterricht genutzt. Dank der finanziellen Unterstützung der Universität Zürich sowie anderer Geldgeber konnte das Programm über die Jahre hinweg kontinuierlich weiterentwickelt werden. Gegenwärtig diskutieren die Verantwortlichen vor allem Strategien zur sinnvollen Integration von Web-2.0-Technologien in *Ad fontes*, damit die etablierten Einheiten zum Selbststudium um Möglichkeiten zum kollaborativen Arbeiten ergänzt werden können. Im Zuge dieser Überlegungen wurde 2009 ein Wiki entwickelt, dank welchem die Studierenden im Rahmen von Lehrveranstaltungen gemeinsam einen Quellenbestand transkribieren und diskutieren können. Ein zentrales Lernziel dieses Pilotprojekts war nebst der Vertiefung paläographischer Fertigkeiten die Heranführung an Techniken und Standards des wissenschaftlichen Edierens historischer Quellen. Obwohl Wikis nur eine mögliche Form kollaborativer Lehr- und Lernformen darstellen, eröffnen die im Rahmen des Projekts gesammelten Erfahrungen interessante Rückschlüsse auf Anforderungen und Potentiale, aber auch auf Probleme und Spezifika bei der Nutzung von Web-2.0-Technologien in den historischen Grundwissenschaften speziell und im universitären E-Learning allgemein.

Abstract

More than seven years ago, the e-learning programme *Ad fontes* went online. Its basic idea has always been to introduce students of history to the challenging work at archives and to improve their palaeographic skills by means of interactive transcribing exercises. By now, *Ad fontes* is regarded as one of the leading tools used as part of a blended learning approach in various practical courses, not only at history departments. After a first successful stage of consolidation, *Ad fontes* now has to face new challenges in the thriving market of Web 2.0 technologies. This calls for a shift from older,

instruction-based training units to more open forms of learning and teaching which soften strict hierarchic roles and promote collaborative user- and authorship. At the University of Zurich, this development led to the idea of introducing a wiki that is on the one hand an extension of *Ad fontes* and on the other hand a tool for practicing editing skills. The wiki is only one possible form of collaborative learning, but it exemplifies and illustrates strategies to deal with new interactive requirements. Apart from a general discussion of e-learning in the first decade of the 21st century and possible implementation of Web 2.0 applications, the focus of this chapter is also on a prospective and rather self-critical reflection of e-learning strategies as such.

1. E-Learning – nach dem Boom die Krise?

Eine Einführung in die Findmittel des Staatsarchivs Zürich, eine interaktive Übung zur päpstlichen Diplomatie, ein Überblick zur altwestnordischen Schriftgeschichte oder die Transkription eines Schüleraufsatzes aus dem Jahr 1914 – dies sind einige Beispiele neuer Lerneinheiten, die in den letzten Monaten bei *Ad fontes* aufgeschaltet worden sind. Das 2001 begonnene Projekt »*Ad fontes* – Eine Einführung in den Umgang mit Quellen im Archiv« gehört unter den webbasierten E-Learning-Programmen im Bereich der historischen Grundwissenschaften wohl zu den etabliertesten Angeboten. Es ist als sogenanntes primäres Lernobjekt ausgestaltet worden, das heisst es stellt ein didaktisch homogenes, inhaltlich in sich geschlossenes virtuelles Gefäss zur zeit- und ortsunabhängigen Selbstbildung dar (Schmale 46–52, 124–131).¹

Doch sind solche primären Lernobjekte und speziell Ausbauprojekte wie die eben erwähnten Lerneinheiten in Zeiten von Web 2.0 eigentlich noch zeitgemäss? Welches Potential wohnt neuen technischen Möglichkeiten inne, die es den Lernenden erlauben, aus einer Fülle von handschriftlichen Quellen die ihnen interessant scheinenden auszuwählen und gemeinsam mit Kolleginnen und Kollegen im virtuellen Raum zu transkribieren? Wie lässt sich die nachhaltige Nutzung von Online-Trainings zur Schulung paläographischer Fähigkeiten garantieren – durch eine engere Anbindung an die Horte handschriftlicher Kostbarkeiten, also die Archive? Unter welchen Bedingungen können und sollen sich Universitäten die hohen Initialinvestitionen sowie die kontinuierlichen Pflegemassnahmen für webbasierte Lernangebote leisten? Und zu guter Letzt: Befindet sich E-Learning nach dem grossen Boom in der Krise?

¹ Schmale, Gasteiner, Krameritsch und Romberg unterscheiden primäre von sekundären und tertiären Lernobjekten. Im Gegensatz zu den strukturell geschlossenen, statische Lernziele verfolgenden primären Lernobjekten legen sekundäre und tertiäre den Fokus auf die Vermittlung von Medienkompetenz, listen ansonsten aber keine fixen Lernziele auf, sondern integrieren mittels kollaborativem Arbeiten im Rahmen einer Veranstaltung auf flexible Weise mannigfaltige Webressourcen in die Lehre. Ein Beispiel dafür wäre das Projekt Deutsch-französische Materialien (DeuFraMat) für den Geschichts- und Geographieunterricht.

Dieser Artikel versucht Lösungen für die aufgeworfenen Fragen anzudenken. Auf der Suche nach Antworten geht der Text von allgemeingültigen Entwicklungen im Bereich des universitären E-Learnings aus, reflektiert diese sodann vor dem Hintergrund des langjährigen Erfahrungshorizonts, den das Projekt Ad fontes durch sein – gemessen am Zeitverständnis des sich rasant verändernden Mediums Internet – nahezu biblisches Alter mitbringt, und versucht letztlich verallgemeinerbare, doch spezifisch auf die Disziplin der historischen Grundwissenschaften gemünzte Schlüsse aus diesen Erfahrungen zu ziehen.

2. Aktuelle Entwicklungen und Kernthesen zum universitären E-Learning

Über sämtliche Disziplinen hinweg wurde E-Learning² gegen Ende der 1990er Jahre zu einem Schlagwort, von dem man sich damals eine Revolutionierung universitärer Lehre im Sinne einer drastischen Reduktion der für ein Studium notwendigen Präsenzzeit in Hörsälen oder Seminarräumen versprach. Derartige Szenarien haben sich in der Zwischenzeit zwar kaum bewahrheitet; vielmehr ist die Anfangseuphorie in weiten Kreisen einem nüchternen Pragmatismus gewichen (Nikolopoulos 56f.). Und doch kann man mit Fug und Recht behaupten, dass wohl für alle, die in der akademischen Lehre tätig sind, der Schritt zurück in eine Zeit ohne IT-gestützte Wissensvermittlung schlechterdings undenkbar scheint. Virtuelle Seminarsitzungen, bei denen dank Software wie Adobe Connect Studierende aus unterschiedlichen Erdteilen gemeinsam einen Text diskutieren, oder Vorlesungen, die infolge überfüllter Räume auch als Podcast zum Download angeboten werden, sind im Lehrveranstaltungsangebot noch immer aussergewöhnlich. Doch mannigfaltige Formen des Blended-Learnings gehören heutzutage wie ganz selbstverständlich zum universitären Alltag, sei es durch Gebrauch von Learning Management Systeme wie moodle, .LRN, ILIAS oder OLAT, sei es durch Miteinbezug spezifischer Selbstlernprogramme aus dem jeweiligen Fachgebiet.

Obwohl also der vor zehn Jahren verbreitete Enthusiasmus über die neuen Möglichkeiten des E-Learnings gegenwärtig nur noch selten anzutreffen ist, ist bei genauem Hinsehen in diesem Entwicklungszweig rein quantitativ mehr im Tun als noch in der visionären Anfangszeit. Diese Feststellung gilt auch für den Bereich der historischen Grundwissenschaften, wie entsprechende Beiträge im ersten Sammelband zum Thema »Kodikologie und Paläographie im digitalen Zeitalter« bewiesen haben (Kamp sowie Cartelli und Palma).³ Es finden pausenlos Veränderungen statt, denen sich

² Schon in den Anfangszeiten wurde der Begriff »E-Learning« allerdings uneinheitlich verstanden, vgl. zu dieser Definitionsproblematik Nikolopoulos, S. 33–35.

³ Vgl. zu diesem Thema auch den von Hiram Kümper geplanten Sammelband zum Thema E-Learning und Mediävistik, welcher voraussichtlich im Herbst 2010 erscheinen wird.

bestehende E-Learning-Tools anzupassen haben und auf die neue Projekte Rücksicht nehmen müssen, wenn sie finanziell gefördert und in der praktischen Nutzung breit akzeptiert werden wollen. Blickt man auf die letzten drei bis vier Jahre zurück, so verdienen es vor allem drei Entwicklungen, explizit hervorgehoben zu werden.

1. Der Kampf um Nachhaltigkeit: Das Bewusstsein für die Notwendigkeit einer langfristigen Planung ist in den Kreisen von E-Learning-Verantwortlichen gewachsen und kann unterdessen mit empirischen Argumenten untermauert werden (Gutbrod, Haug und Wedekind). Ohne eine Zusicherung für die langfristige finanzielle Förderung der technischen und inhaltlichen Pflege eines E-Learning-Tools machen die hohen Anfangsinvestitionen zur Entwicklung des technischen Frameworks und der inhaltlichen Module wenig Sinn. Dem steht die Erfahrung gegenüber, dass es für die Geldgeber sowohl wegen des abschätzbaren Kostenrahmens wie auch aus Prestige Gründen attraktiver ist, den Aufbau neuer Projekte, nicht aber deren spätere Pflege zu fördern. Für die Universitäten, welche diese Aufgabe noch am ehesten zu übernehmen gewillt sind, ist es gerade in Zeiten von Knappheit öffentlicher Finanzen verführerisch und bis zu einem gewissen Grad auch verständlich, in jenen Bereichen zu sparen, wo dies die strukturellen Gegebenheiten erlauben; leider fallen E-Learning-Engagements naheliegenderweise in diese Kategorie. So erweist es sich oft als schwierig, die personellen und monetären Ressourcen sicher zu stellen, um als E-Learning-Projekt in einem sich schnell wandelnden Umfeld auf Dauer bestehen und innovative Neuentwicklungen präsentieren zu können.
2. E-Learning als E-Research: Die Akzeptanz von Formen des webbasierten Lernens steigt an Universitäten gemeinhin dann an, wenn dieselben nebst dem individuellen Wissenserwerb zusätzlich entweder unkomplizierte, weil standardisierte Formen der Leistungsermittlung erlauben (E-Assessment) oder aber wenn durch sie Querverbindungen zu aktuellen Forschungstrends geschaffen werden (E-Research). Die Studierenden sollen mit Hilfe von E-Learning also den eigenen Forschungstrieb entdecken können und im Idealfall sogar konkrete Beiträge zu einem Forschungsvorhaben leisten (Henri).
3. Der Ruf nach Offenheit, Flexibilität und kollaborativen Lernformen: In einer Zeit, da Studierende an die unter dem Begriff »Web 2.0« subsumierten Angebote wie Facebook, Wikipedia oder Blogs gewöhnt sind, erwarten sie als Nutzerinnen und Nutzer von E-Learning-Tools gleichfalls interaktive Schnittstellen und Möglichkeiten zur Mitgestaltung sowie zum Informationsaustausch. Auch didaktisch wird diese Entwicklung begrüsst – kollaborative Lernformen, bei denen Kenntnisse und Fertigkeiten nicht mittels eines von Experten vorgespurten, auf Nachvollzug ausgelegten Lernarrangements vermittelt, sondern schwergewichtig im Rahmen gegenseitiger Interaktion und Kommunikation unter den Lernenden erworben werden, tragen den heute verbreiteten konstruktivistischen Wissensvorstellungen

eher Rechnung (Mandl und Krause 5, 11–13). Entsprechende Lernumgebungen zeichnen sich durch eine Reihe von Merkmalen aus; vor allem dem Lernen an komplexen, lebensnahen und ganzheitlichen Problemkreisen, der arbeitsteiligen Selbstorganisation in einer sozialen Gruppe sowie der Möglichkeit, bereits erworbene Fähigkeiten oder Kenntnisse sachdienlich einzubringen, werden gesteigerte Bedeutung zugemessen. Auf diese Weise verstandenes Lernen wird im Gegensatz zum instruktionalistischen Unterricht als aktiver, personalisierter und dialogischer Prozess erlebt, bei dem Fehler nicht nur erlaubt, sondern erwünscht sind, weil das Erkennen und die Korrektur derselben zu bleibenden Erfahrungen und somit zu einer pragmatischen Verinnerlichung neuen Wissens führt (Dubs). Diese methodischen Grundsätze der konstruktivistischen und dialogischen Didaktik sind keineswegs an den klassischen Unterricht gebunden, sondern haben im Bereich des E-Learnings nicht minder Gültigkeit (Ruf 192–196).

Das folgende Hauptkapitel vertieft und exemplifiziert die drei skizzierten Strukturveränderungen, wobei ein besonderer Fokus auf den dritten Punkt, das kollaborative E-Learning zur Verbesserung paläographischer Fähigkeiten, gelegt werden soll. In diesem Zusammenhang werden erste Erfahrungen mit einem noch nicht öffentlich zugänglichen, speziell für quellenkundliche Lehrveranstaltungen konzipierten Wiki zum gemeinschaftlichen Edieren von Handschriften vorgestellt. Dieses Wiki erweitert das bestehende Angebot von Ad fontes, indem es für einmal keine weitere Lerneinheit, sondern eine neue, flexibel nutzbare Lernform hinzufügt. Zum Schluss jedes Teilkapitels wird im Sinne einer Bilanz versucht, die inneren Entwicklungen bei Ad fontes in den grösseren Kontext der E-Learning-Landschaft im Bereich »Paläographie und Archivistik« einzubetten.

3. Persistenz und Wandel – der schmale Grat zwischen bleibenden Werten und sinnvoller Anpassung

3.1. E-Learning und der Kampf um Nachhaltigkeit

Nicht selten wird bei E-Learning-Projekten die Notwendigkeit langfristiger Planung und die Sicherstellung einer nachhaltigen finanziellen Förderung unterschätzt (siehe oben, 2). Das Einwerben von Geldern für die zeitlich befristete Durchführung eines bestimmten Projekts unterliegt freilich anderen Gesetzmässigkeiten als die unmittelbar an die eigentliche Umsetzungsphase anschliessenden Anstrengungen um Zusicherung einer dauerhaften Sockelfinanzierung. Doch jede bereits erstellte Lerneinheit braucht sowohl inhaltliche Pflege, wenn sie den Anspruch auf wissenschaftliche Aktualität ernst nehmen will, als auch technische Wartung, wenn sie mit den in ständigem Fluss befindlichen Soft- und Hardwareanforderungen Schritt halten will.

So sind auch in der Geschichte von Ad fontes zwei Etappen des inhaltlichen Ausbaus zu unterscheiden, nämlich erstens die initiale Aufbauphase, wo ein mehrköpfiges Team am Historischen Seminar die Kerninhalte des Programms entwickelte. Viele Lerneinheiten aus dem Bereich der historischen Grundwissenschaften und der Archivkunde stammen aus dieser dynamischen Zeit, in der nicht nur von der Universität Zürich, sondern auch von Stiftungen grosse Summen investiert worden sind (Kränzle und Ritter 194–196). Daran schloss nach einigen Jahren zweitens eine Konsolidierungs- und Ausbauphase an, die bis heute anhält. Seit 2006 finanziert sich Ad fontes aus einem Grundsicherungsbeitrag der Philosophischen Fakultät und aus Geldern der von der Universität Zürich alljährlich ausgeschriebenen »Initiative Interaktives Lernen«, mit der befristete Ausbauprojekte gefördert werden. Dadurch ist die Sicherstellung von Pflege und Aktualisierung des Bestehenden seit fünf Jahren eng an einen sanften inhaltlichen Ausbau gekoppelt. Erfreulich an dieser förderpolitisch bedingten Verzahnung war, dass sich Ad fontes inneruniversitär immer mehr gegenüber anderen Disziplinen öffnete, darunter die deutsche Sprachwissenschaft, das Mittellatein, die Nordistik (Altnordisch) und die Romanistik (Altfranzösisch). Die zunehmende Interdisziplinarität macht insofern Sinn, als archivistische und paläographische Kenntnisse in allen historisch orientierten Kulturwissenschaften von hohem Nutzen sind und weil das Kosten-Nutzen-Verhältnis der entwickelten Lerneinheiten durch das gemeinsame technische Fundament steigt. Zugleich erfordert die Zusammenarbeit trotz des vereinenden paläographischen Grundinteresses natürlich immer auch Kompromisse. So mussten beispielsweise die in der Nordistik sonst üblichen Transkriptionsregeln recht stark im Sinne einer normalisierten, Ligaturen auflösenden statt buchstabengetreuen Übertragung angepasst werden, weil die für die ursprünglich angestrebte originalgetreue Wiedergabe nötigen Sonderzeichen in den einem webbasierten E-Learning-Programm zur Verfügung stehenden Standardschriftsätzen schlichtweg fehlten.⁴ Damit die für altnordische Transkriptionsübungen unabdingbaren Zeichen Eth (Ð, ð) und Thorn (Þ, þ) gleichwohl eingegeben werden können, musste eigens eine entsprechende Eingabehilfe programmiert werden.

Als wichtig bei der Wahl neuer universitärer Kooperationspartner entpuppten sich vorangehende, in der konventionellen Lehre gesammelte Erfahrungen mit den für Online-Transkriptionsübungen ausgewählten Quellenstücken. Die Qualität einer Transkriptionsübung bemisst sich zu einem wesentlichen Grad an der Präzision der bei Bedarf zu jedem Wort abrufbaren Tipps. Das Verfassen derselben erfordert mithin ein zuverlässiges Antizipieren jener Probleme, mit denen Anfänger zu kämpfen haben. Nicht selten sind die Produzenten der Übungen, für welche das Lesen eines bestimmten Schrifttyps zur Selbstverständlichkeit geworden ist, dazu nicht mehr in der Lage;

⁴ Vgl. zur allgemein unbefriedigenden Integration historischer Glyphen in die verbreiteten Unicode-Zeichensätze die Website der Medieval Unicode Font Initiative.

es sei denn, sie kennen die typischen Stolpersteine einer Quelle dank mehrmaliger Verwendung derselben in der Lehre. Nach wie vor sind solche manuell verfassten Feedbacks der bei Ad fontes vor einigen Jahren eingeführten Technik automatisch generierter Tipps, die den ersten falschen Buchstaben eines Worts angeben und das Wort bei Bedarf gleichsam buchstabieren, didaktisch überlegen. Praktisch jedes ernst zu nehmende E-Learning-Angebot für paläographisches Lesetraining bietet denn auch die Möglichkeit, bei Bedarf ausgewählte Hilfestellungen einzublenden, so z. B. Paläographie Online, das paläographische Lesetraining für lateinische Schriften des 5.–20. Jahrhunderts von Thomas Frenz oder die Tutorials der National Archives.

Trotz der Vorteile einer mit den Jahren immer breiter werdenden interdisziplinären Abstützung, die allerdings als Negativum auch einer zunehmenden Unübersichtlichkeit Vorschub leistete, wirft die Politik, sich über Ausbauprojekte zu finanzieren, auch Probleme auf. Besonders schwierig ist es, rein technisch motivierte Neuerungen zu finanzieren. Eine solche steht allerdings gegenwärtig bei Ad fontes an. Bis vor kurzem basierten alle Übungen dieses Programms auf der Shockwave-Technologie. Bei der Neuentwicklung von Ad fontes im Jahr 2001 war dieses von Macromedia entwickelte Plug-in die beste Wahl zur Umsetzung interaktiver Aufgaben innerhalb von HTML-Seiten. Die Situation hat sich unterdessen durch zwei Entwicklungen entscheidend gewandelt. Erstens wird Shockwave nach der Übernahme durch Adobe nur noch halbherzig weiter entwickelt, so dass Ad fontes mit immer mehr technischen Inkompatibilitäten zu kämpfen hatte, was wiederum zu einem latenten Missmut bei den Studierenden und zu einem erhöhten Betreuungsaufwand führte. Zweitens gibt es heutzutage vielfältige Möglichkeiten, interaktive Inhalte auf Webseiten mittels AJAX (Asynchronous JavaScript and XML) zu implementieren. Dass diese Technologie genügend umfangreiche Funktionalitäten bereitstellt, um auch in Transkriptionsübungen Verwendung zu finden, beweisen schon seit mehreren Jahren die Seiten von Paläographie Online, wo die Nutzerinnen und Nutzer zwischen einer Shockwave- und einer Javascript-basierten Version auswählen können. Obwohl die Dringlichkeit dieser technischen Umstellung bereits länger bekannt war, konnte sie vor einigen Monaten nur deshalb in Angriff genommen werden, weil Ad fontes im Förderverein⁵ ein Finanzierungsinstrument zur Seite steht, das gewillt war, die Kosten für ein derartiges Projekt zu übernehmen, das zwar für die Nachhaltigkeit des Angebots absolut essentiell ist, das aber vor dem Hintergrund, dass das Programm danach nicht im eigentlichen Sinn mehr bietet, auf den ersten Blick leider wenig prestigeträchtig anmutet.

Dass sich die Sicherstellung einer stabilen Finanzierung zunehmend schwierig gestaltet, liegt natürlich auch daran, dass gegenwärtig niemand mehr glaubt, E-Learning sei die kostengünstigere Variante konventionellen Unterrichts. Wer also

⁵ Dieser Förderverein rekrutiert sich aus Nutzerinnen und Nutzern, die Ad fontes jedes Jahr freiwillig mit einem finanziellen Beitrag unterstützen.

heute Unterstützung für ein Projekt einfordert, muss argumentativ überzeugend darlegen, welchen pädagogischen Zusatznutzen dieses abwirft (Seufert und Euler 57). Ein solcher, soweit eine erste Zwischenbilanz, ist im Fach Geschichte vor allem dann vorhanden, wenn die aufbereiteten Lernszenarien grundlegend sind, wenn sie also möglichst stark auf die fachliche Methodologie und wissenschaftliche Kernkompetenzen fokussieren und nur so viel wie nötig auf Inhalte beziehungsweise konkrete historische Themen eingehen.⁶ Dies sichert ihre langfristige Aktualität und damit ihre adaptive Verwendbarkeit in der klassischen Lehre. Die Schulung von Kompetenzen im Bereich der Quellenkritik und der historischen Grundwissenschaften vermag diese Anforderung in hohem Masse zu erfüllen.

3.2. E-Learning als Brücke zur Forschung

Die erfolgreiche Verbindung von Forschung und Lehre gilt als konstitutives Merkmal einer Universität. Nachdem in den letzten zwei Jahrzehnten in beiden Feldern die Bedeutung IT-basierter Lösungen zugenommen hat, liegt es nahe, verstärkt über mögliche Schnittstellen zwischen E-Learning und E-Research nachzudenken, damit auch im Netzzeitalter die traditionell enge Koppelung der akademischen Ausbildung an die wissenschaftlichen Trends aufrechterhalten werden kann (vgl. dazu die Podcasts des E-Learning-Centers der Universität Zürich).

Da historische Forschung zu einem wesentlichen Anteil in Archiven stattfindet, wo gegenwärtig die Umstellung auf computergestützte Technologien in diversen Bereichen in vollem Gang ist, wurde das Potential eines lockeren Zusammengehens öffentlicher Archive mit Ad fontes als etabliertem E-Learning-Programm bald augenfällig. Einen ersten wichtigen Partner zur Verwirklichung dieser Strategie fand Ad fontes im Staatsarchiv Zürich, dessen Quelleneditionsstrategie in Zukunft vielfältige webgestützte Forschungsmöglichkeiten bieten wird.⁷ Da es mit dem reinen Online-Zugriff auf Quellen indes noch nicht getan ist, sondern auch die für die jeweilige Quellenart spezifische Methodik des Suchens und Auswertens trainiert werden muss, wurden gemeinsam mit Archivmitarbeitenden E-Learning-Einheiten zum Umgang mit Verwaltungsquellen des 19. Jahrhunderts konzipiert, die im Laufe dieses Jahres aufgeschaltet werden sollen. Diese Übungen sollen nicht nur die unbestritten wachsende Relevanz von Online-

⁶ Vgl. dazu die Äusserung von Manfred Thaller an den 2. Schweizerischen Geschichtstagen, die als Audiofile zugänglich ist.

⁷ Aktuell führt dieses Archiv vier diesbezügliche Grossprojekte durch, nämlich die Digitalisierung und Publikation historischer Karten und Pläne (2008–2014), die Transkription und Digitalisierung der Regierungsratsbeschlüsse und Kantonsratsprotokolle des Kantons Zürich seit 1803 (2009–2015), die Erstellung einer »Ehe-Datenbank« für das 16. bis 18. Jahrhundert (2008–2012) sowie eine umfängliche digitale Rechtsquellen-Edition für den Kanton Zürich (2010–2017). Darüber hinaus wirkt es federführend an der Etablierung eines Suchportals (Archives Online) für den Zugriff auf Online-Datenbanken öffentlicher Archive mit.

Recherchen für die Archivarbeit beleuchten, sondern auch ein zentrales Augenmerk auf die Entkräftung der Vorstellung legen, dank der Fülle an bequem im Internet verfügbaren, bereits transkribierten Ressourcen seien profunde paläographische und archivistische Kenntnisse gleichsam unnötig geworden.

Für im Bereich Paläographie und Archivistik beheimatete E-Learning-Projekte schlummert also unseres Erachtens noch ein grosses, ungenutztes Potential in einer direkteren Anbindung an ausgewählte Archive und an langfristig angelegte Forschungs- sowie Editionsprojekte. Solche Kooperationen bringen einen beidseitigen Gewinn, zumal sich das E-Learning-Projekt ein neues, über die universitäre Stammklientel hinausgehendes Zielpublikum erschliesst, während die Archive Anschluss an eine Informations- beziehungsweise Lernplattform erhalten, deren Möglichkeiten jene eines konventionellen Webauftritts mit seinem eindimensionalen Informationsfluss übersteigen. Zudem senkt die Verfügbarkeit interaktiver Trainingsmöglichkeiten zur Schulung des historischen Handwerks die verbreitete Schwellenangst vor dem Umgang mit handschriftlichem Material.

3.3. Von der Instruktion zur Kollaboration – edieren lernen mit dem Ad fontes-Wiki

E-Learning, so wird gemeinhin argumentiert, bringt den Vorteil mit, dass in der virtuellen, netzartig angelegten Lernumgebung die Lösungswege und die Zeit, die man für eine Aufgabe benötigt, individuell gewählt werden können (Flindt 51). Via Interaktion mit der flexiblen technischen Architektur eines Programms findet eine im Gegensatz zum klassischen Präsenzunterricht weitgehend selbstbestimmte Auseinandersetzung mit den Inhalten statt, die subjektiven Präferenzen ebenso Rechnung trägt wie sie den Lernenden allfällige stoffliche Defizite vor Augen führt und Wege zur Behebung derselben aufzeigt. Trotz dieses Vorteils haben E-Learning-Angebote mit interaktiven Übungen, bei denen Eingaben mit einer Musterlösung abgeglichen und Rückmeldungen durch Auswahl aus einer feststehenden Sammlung von Tipps erfolgen, den unübersehbaren Nachteil, dass ihre instruktionalistische Grundstruktur der im nichtschulischen Alltag gepflegten Feedback-Kultur beim Lernen nicht entspricht und dass insbesondere der Interaktion unter den Lernenden zu wenig Beachtung geschenkt wird. Darüber hinaus bleiben sie auf den gerade in den Kulturwissenschaften engen Bereich eindeutig entscheidbarer Lernszenarien mit Richtig-Falsch-Schema beschränkt. Um dieses Manko zu beheben, so die aktuellen Theorien im Bereich des universitären E-Learnings (Zimmermann 180f.), müssen die klassischen Selbstlerneinheiten um Lernarrangements erweitert werden, die das gemeinsame Arbeiten an einer bestimmten Aufgabe erlauben und in denen die Studierenden wechselseitig vom Wissenshorizont der anderen profitieren können. Dass hier Blended-Learning als Konzept die gesuchte Kombination von traditioneller Präsenzveranstaltung mit einem kollaborativ oder individuell angelegten E-Learning-

Angebot bietet und Mehrwerte generieren kann, wurde in der Fachliteratur bereits mehrfach hervorgehoben (Geldsetzer und Strothmann sowie Mankel). Die Nutzung von Web-2.0-Tools für die Lehr-/Lernumgebung führte sogar zum Begriff »E-Learning 2.0« (Kerres). Ob ein solcher Lösungsansatz im spezifischen Bereich der Paläographie wirklich hält, was die allgemeine Theorie verspricht, wird im nachfolgenden Erfahrungsbericht zu einem noch nicht öffentlich zugänglichen Erweiterungstool von Ad fontes kritisch geprüft.

Bisher fand Ad fontes klassischerweise Einsatz im Rahmen von Proseminaren oder Basismodulen, um ein rudimentäres hilfswissenschaftliches Grundwissen sicherzustellen. Die langfristige Konsolidierung und Erweiterung paläographischer Skills erfordert demgegenüber ein anderes lerntheoretisches Setting (Schmale 47–52). Auch das Vertrautwerden mit Grundlagen der Editionstechnik kann nur schwer mittels interaktiver Übungen vermittelt werden, sondern wird in erster Linie durch selbstorganisiertes Lernen am realen Beispiel geschult. Um als Pilotprojekt im Praxistest eine mögliche Erweiterung von Ad fontes zu evaluieren, wurde deshalb im Sommer 2009 gemeinsam mit dem Ad fontes-Mitbegründer Gerold Ritter ein Wiki-basiertes Editionstool entwickelt, das genau diesen Konnex zwischen präsenzunterrichtbasiertem Wissen und webbasierten Lerneinheiten mit Quellenmaterial schafft. Das passwortgeschützte Editionstool wurde im Rahmen des Seminars »Lokale Dimensionen der territorialen Herrschaft im Spätmittelalter« angelegt und verfolgte als weiteren Hauptzweck die Digitalisierung einer Briefsammlung aus dem Spätmittelalter im Historischen Bürgerarchiv Thun. Es handelt sich um sogenannte Missiven⁸ zwischen der Obrigkeit der Stadt Bern und deren lokalen Herrschaftsvertretern. Diese Sendschreiben stammen aus dem Zeitraum zwischen 1380 und circa 1500, wobei der überwiegende Teil in den Jahren von 1430 bis 1450 entstanden ist. Die dichte Überlieferung von knapp über 2500 Missiven erlaubt einen ausserordentlich aufschlussreichen Einblick in Herrschaft und Administration vor Ort und im Alltag. Dieser Zugriff auf Verwaltungsalltag ermöglicht das Aufgreifen aktueller Diskussionen um »state-building from below«, indem hier soziale Organisationsformen (Klientelismus, Familien- und Verwandtschaftsnetzwerke, Korporationen, etc.) greifbar werden, die in einem Gegensatz zu den in der früheren Historiographie dominierenden Vorstellungen von politischen und verfassungsgeschichtlichen Top-down-Prozessen stehen (Teuscher). Abgesehen vom forschungswissenschaftlichen Gewinn eignen sich die Missiven aber auch aus zwei pädagogisch-didaktischen Gründen. Einerseits handelt es sich bei ihnen um kurze Schreiben mit einer durchschnittlichen Länge von fünf bis zehn Zeilen, das heisst, die Portionierung der einzelnen Schriftstücke muss nicht künstlich vorgenommen werden, sondern ergibt sich aus den Quellen selbst. Andererseits ermöglicht die

⁸ Als Missiven werden Briefe resp. Sendschreiben mit amtlichem Charakter bezeichnet (Teuscher 366f.).

Alltagsnähe des Inhalts einen schnellen Einstieg ins Quellenmaterial, ohne allzu viel Vorwissen abzuverlangen.

Das Projekt verfolgte also mehrere Ziele: Erstens sollten im Rahmen von Seminaren aktuelle Forschungsfragen mit einer quellengestützten Editionsarbeit verbunden werden, zweitens sollte ein Quellenkorpus geschaffen werden, das über mehrere Jahre hinweg genutzt und erweitert werden kann, und drittens sollten die Studierenden die Möglichkeit erhalten, über längere Zeit je nach Bedürfnis allein, im Kollektiv oder unter Anleitung an Transkriptionen respektive kleineren Editionen zu arbeiten und diese Erkenntnisse wiederum in Qualifikationsarbeiten einzubringen.

Die dem Wiki zugrundeliegende Hypertextstruktur erlaubt in Verbindung mit dem intuitiv bedienbaren Content-Management kollaboratives Arbeiten an Texten und geht von der Idee »kollektiver Intelligenz« aus (Iske und Marotzki). Vorteile des Wikis sind neben der als bekannt vorauszusetzenden Benutzeroberfläche die relativ einfache Markup-Sprache, der Wikitext. Der Entscheid, die etablierte Software von MediaWiki als Grundlage des Editionstools zu verwenden, erwies sich in mehrfacher Weise als ideal zur Erfassung, Bearbeitung und Interpretation historischen Quellenmaterials. Das Quellenkorpus besteht aus Digitalisaten der Thuner Missiven sowie Scans der dazugehörigen Regesten und Register, welche mittels OCR (Optical Character Recognition) erfasst worden waren. Jeder Missive wurde eine Seite im Wiki zugeordnet und mit folgenden Kategorien der Bearbeitung versehen: Bild (recto und verso), Datum der Missive, Kurzregest, Transkription, Kommentar und Anmerkungen. Ausserdem erlaubt es das Wiki, Interpretationsansätze und Problemstellungen auf Diskussionsseiten (angelegt pro Missive) zur Debatte zu stellen. Zusätzlich lassen sich unterschiedliche Versionen miteinander abgleichen und jederzeit rückgängig machen, wobei auch die Möglichkeit besteht, endgültige Fassungen zu markieren und schreibgeschützt zu publizieren. Die Zugriffssteuerung für definierte Benutzergruppen ist hier zentral. Als eines der wenigen OpenSource-Angebote in diesem Bereich bringt MediaWiki überdies von Hause aus einen doppelten Anmerkungsapparat und folglich eine für die Editionstätigkeit unabdingbare Funktionalität mit. Die aktuelle Forschungsdiskussion zur digitalen Editorik und insbesondere zu neuen Lösungen und Ansätzen der Textkodierung sowie zu multiplen Editionstextlevels verfolgen weitergehende Ziele.⁹ Als einführendes, auf eine Aneignung traditioneller historisch-kritischer Editionstechniken ausgerichtetes Lerntool erhoben wir bewusst nicht den Anspruch, massgeschneiderte Lösungen bereitzustellen unter Rückgriff auf das heute theoretisch Machbare. Das Wiki bietet durch das Miteinander von digitaler Reproduktion des bildlichen Originals und Transkription des Inhalts die Möglichkeit, neuere methodische Ansätze zur Materialität der Texte wie z. B. die *material philology* aufzunehmen (Nichols). Zudem sind auf dem

⁹ Vgl. aktuelle Forschungsprojekte zu digitaler Editorik im Rahmen des Instituts für Dokumentologie und Editorik <http://www.i-d-e.de/>. Zu erwähnen ist dabei das Editionswerkzeug EditMOM (Burkard), welches kollaborative Urkunden-Erschliessung ermöglicht und fördert.

Wiki Hilfsmittel und Einführungstexte leicht zugänglich in üblichen Textdateien auf der Startseite platzierbar, so dass beispielsweise die Registerseiten mit Volltextsuche nach Orten, Sachen und Personen abgefragt werden können.

Dieses Editionstool wurde im Herbstsemester 2009 erstmals eingesetzt und getestet. Im Verlaufe der Lehrveranstaltung wurden Fertigkeiten vom einfachen Lesen der Quellen bis hin zu editionsfertigen Transkriptionen mit einem zweiteiligen wissenschaftlichen Apparat schrittweise aufgebaut, bevor die Studierenden das erworbene Können im Rahmen einer wissenschaftlichen Studie praktisch unter Beweis stellen mussten. Das Lernsetting gestaltete sich als dreistufiges Angebot von Präsenzunterricht, individuellem Training am eigenen Computer ausserhalb der regulären Veranstaltung und persönlichen Coachingrunden. Diese Kombination (Blended-Learning) ermöglichte es, die Präsenzzeiten als Rahmungen zu konfigurieren, innerhalb derer sowohl inhaltliche Wissenseinheiten wie auch offene Feedbackrunden eingeführt werden konnten. Das selbständige Einüben, Konsolidieren und Erweitern der Kompetenzen im selbständigen Arbeitsteil sowie die damit verknüpften Coachingsitzungen mit den Dozierenden konnten den individuellen Unterschieden bezüglich Wissensstand, Zeitmanagement, Bedürfnissen und Zielsetzungen gerecht werden. Während sich die interaktiv angelegte Arbeit an den Texten im Präsenzunterricht als gewinnbringend erwies, stellte sich die im Selbststudium angelegte Transkriptionsarbeit für die Studierenden zunächst als schwierig dar. Das Wiki vermochte den Sprung vom vermittelten Wissen hin zur selbständigen Transkription nur bedingt zu überbrücken. Obwohl es die Möglichkeit bietet, auf den Diskussionsseiten gezielt zu einer Missive Hilfestellungen oder Vorschläge sowohl sprachlicher wie inhaltlicher Natur einzuholen, zeigte sich das Tool in der Regel nur dann als ergiebig, wenn nicht Peer-Reviews stattfanden, sondern sich erfahrener Studierende respektive Tutoren und Tutorinnen darin einbrachten. Ist der Erfahrungsschatz der Bearbeitenden also in etwa vergleichbar, stellt sich der Nutzen solcher Diskussionen nur bedingt ein. Eine entscheidende didaktische Erkenntnis war denn auch, dass dem Faktor Zeit bei der Einübung und Konsolidierung von Transkriptionswissen und -strategien viel Bedeutung beigemessen werden muss. Zudem sollte die Möglichkeit eines situationsgerechten Coachings gegeben sein, um auf Probleme eingehen zu können. Was diese Probleme anbelangt, so bestand der Grossteil vor allem in Unsicherheiten bei Lesungen oder Abkürzungsaufösungen. Das Bedürfnis nach autoritativer Bestätigung einer Lesart oder einer ganzen Transkription wurde von vielen Studierenden artikuliert und muss als solches letztlich auch gewährleistet werden. Obwohl eine der Grundideen in Wikis die Aufweichung von Rollenverteilungen und hierarchischen Lehrstrukturen ist (Kerres 6f.), scheint doch der attestierte Leistungsausweis einer »korrekten« Transkription für die meisten Studierenden nach wie vor zentral zu sein. Die Diskussionsseiten brachten aber dahingehend wichtige Einsichten, dass Probleme zwar nicht unbedingt gelöst, aber auf einer Metaebene diskutiert und geteilt werden konnten. Dadurch konnten von den Dozierenden sowohl

kollektive wie auch individuelle Lehrprozesse verfolgt und allfällige Schwierigkeiten oder Unsicherheiten im Präsenzunterricht aufgenommen und thematisiert werden.

Die ersten Transkriptionsarbeiten waren zwar als Einstiegsübungen angelegt, jedoch war ganz allgemein die Hemmschwelle der Studierenden – trotz »Generation Facebook« – etwas Unfertiges, womöglich auch »Falsches« im Web zu publizieren, sehr hoch.¹⁰ Diesem Effekt konnte entgegengewirkt werden, indem einerseits die Problematik in der Lehrveranstaltung selbst diskutiert und andererseits eine allgemeine Gewöhnungsphase an die neue Lehrform hin zu deren Akzeptanz als Notwendigkeit angenommen wurde.

Das vorläufige Resümée im Sinne verallgemeinerter Perspektiven zur Nutzung kollaborativer E-Learning-Tools im Bereich der Paläographie fällt wohlwollend, doch keineswegs uneingeschränkt positiv aus: Ein relativ unkompliziert implementierbares Editionstool wie das eben vorgestellte leistet einen wertvollen, ergänzenden und auf individuelle Lernprozesse eingehenden Beitrag beim Erwerb paläographischer und kodikologischer Skills. Erst die aufeinander abgestimmte Kombination von Präsenzunterricht, klassischen Selbstlerneinheiten auf Ad fontes und kollaborativem Arbeiten an Transkriptionen im Wiki generierten aber den erhofften Zusatznutzen, der eben nicht nur den Wissenserwerb, sondern auch dessen Konsolidierung ins Zentrum stellt. Ganz grundsätzlich lässt sich festhalten, dass die Lernprozesse der Studierenden mit solchen kombinierten Angeboten transparenter gemacht werden und dass durch kollaborative Tools spezifische Transkriptionsprobleme zwar nur selten im gegenseitigen Austausch unter den Studierenden gelöst, wohl aber von Dozierenden erkannt und behoben werden können.

4. Bilanz – ohne zeitliche und finanzielle Investitionen kein didaktischer Gewinn

E-Learning im Gebiet der Paläographie – das haben die drei diskutierten Entwicklungstrends gezeigt – bewegt sich in einem spannungsreichen Feld technischer Veränderungen, monetärer Vorgaben, potentieller Kooperationen, didaktischer Theorien und studentischer Ansprüche. Auch wenn sich auf historische Grundwissenschaften hin orientierte Online-Angebote wie Ad fontes einer Gesellschaft, in welcher der webbasierte Austausch von Informationen alltäglich ist, einerseits unbestritten anpassen müssen, dürfen sie andererseits aus finanziellen und didaktischen Überlegungen nicht blind jedem Trend folgen. Obwohl die »Generation Facebook« der Studierenden in technischer Hinsicht keinerlei Hemmschwellen vor Web-2.0-Anwendungen hat und solche Funktionalität sogar explizit oder implizit einfordert, tut sie sich mit

¹⁰ Bei der privaten Nutzung von Web 2.0 stehen in erster Linie soziale Netzwerke im Zentrum, und diese werden vor allem zur Kommunikation und zum »Alltagsmanagement von Freundschaften« verwendet. Relativierend und kritisch zum Begriff und Phänomen »Net-Generation« äussert sich Schulmeister 129f.

den Ansprüchen, die einem mit wissenschaftlichen Ansprüchen daherkommenden webbasierten Informationsaustausch inhärent sind, schwer.¹¹ Die Einführung einer technischen Infrastruktur zum kollaborativen Arbeiten an Handschriften macht also zwar durchaus Sinn, bringt aber unseres Erachtens nur dann etwas, wenn sie nicht freiwillig und unverbindlich zur Verfügung steht, sondern institutionalisiert genutzt wird, das heißt im Rahmen von Lehrveranstaltungen mit konkreten Lernzielen und einem gemeinsamen Erkenntnishorizont. Falls diese Voraussetzung gegeben ist, lässt sich die notwendige Nutzungsintensität und mithin auch ein pädagogischer Profit erreichen. Doch selbst wenn der zur Mitarbeit verpflichtende Rahmen einer auf handschriftliche Quellen fokussierten Lehrveranstaltung mit der technischen Möglichkeit eines auf die spezifischen Bedürfnisse hin optimierten Editionswikis dies alles sicherstellt, erfordert dergestalt eingebettetes E-Learning einen hohen individuellen Betreuungsaufwand. Die kollaborative Arbeit mit Handschriften kämpft hier auch mit einer erschwerenden Besonderheit: Obwohl es für jeden Schrifttypus bekannte Hürden und Schwierigkeiten gibt wie zum Beispiel das cc-a der Halbunziale oder die Mehrdeutigkeit der winkligen deutschen Kurrentschrift, entsteht ein Grossteil der Lese Probleme in sehr subjektiven Kontexten der Auseinandersetzung zwischen einer Person und einer bestimmten Quelle. Damit eine andere Person Hilfe leisten kann, muss sie verhältnismäßig viel Zeit investieren, bis sie den inhaltlichen Kontext der Quelle grob erschlossen und die Spezifika der jeweiligen Hand erkannt hat. Diesen hohen Einsatz leisten andere Studierende in der Regel nicht aus bloßer Solidarität, sondern eben nur dann, wenn sie ein Eigeninteresse am Entziffern derselben Quellenstelle mitbringen.

Universitäres E-Learning befindet sich, um auf die eingangs aufgeworfenen Fragen zurückzukommen, also nicht in einer eigentlichen Krise, wohl aber in einem Prozess der differenzierteren Selbstwahrnehmung, der kritischen Selbstfindung und der wachsenden Überlappung mit anderen universitären Arbeitsgebieten wie E-Assessment oder E-Research. Das in diesem Artikel vorgestellte Editionswiki stellt den Versuch dar, die vorerst noch begrifflich und institutionell getrennten Sphären auch in der Disziplin »Paläographie« einander anzunähern. Dass Anstrengungen solcher Art didaktisches Gespür, zeitlichen Aufwand und finanzielle Ressourcen verlangen, darf nicht entmutigen, vermag doch der synthetisierende Rückgriff auf unterschiedliche Lehr-/Lernwelten unter optimalen Bedingungen sehr wohl einen schätzenswerten Mehrwert zu generieren. Im Falle der im letzten Kapitel geschilderten Erfahrungen gehörten zu diesen Bedingungen auch die konventionellen Tutorien und Trainings von Ad fontes, die also, um eine weitere anfangs gestellte Frage zu beantworten, ihre Daseinsberechtigung alles andere als verloren haben. Gerade weil Studierende, die beim Transkribieren der im Wiki vorhandenen Missivensammlung Mühe bekundeten, das fehlende paläographische

¹¹ Eine differenzierte Einschätzung des Potentials von Web-2.0-Technologien für die universitäre Lehre gibt Reinmann 15f.

Rüstzeug durch Nutzung der klassischen interaktiven Einführungen und Transkriptionsübungen bei Ad fontes unkompliziert sowie ohne Gesichtverlust aufarbeiten konnten, machte die Arbeit mit der neuen technischen Möglichkeit Sinn. Ebenfalls als Schlüssel zum Erfolg erwies sich die Koppelung an konventionelle Beratungsangebote, also zusätzliche Sprechstunden, in denen Leseprobleme mit Expertinnen und Experten diskutiert werden konnten.

Ad fontes wird den Weg der sanften Innovation und des kontinuierlichen Ausbaus weitergehen. Festhalten wird es an seiner Politik der Fokussierung auf methodische Grundfertigkeiten und exemplarische Lernszenarien, zumal sich in der naturgemäss kleinen Nische des paläographischen E-Learnings nur mit einer strategischen Ausrichtung auf den Erwerb verallgemeinerbarer Fähigkeiten ein nachhaltiger Gegenwert für Investitionen sicherstellen lässt.

Bibliographie

- Ad fontes. Eine Einführung in den Umgang mit Quellen im Archiv.* Zürich: Universität Zürich, 2001–2010. <<http://www.adfontes.uzh.ch/1000.php>>.
- Archives Online.* Staatsarchive der Kantone Zürich, Thurgau, Zug und Basel-Stadt – Archiv für Zeitgeschichte, 2010. <<http://www.archives-online.org/>>
- Burkard, Benjamin. »EditMom – ein spezialisiertes Werkzeug zur kollaborativen Urkunden-Erschliessung.« *Digitale Diplomatie. Neue Technologien in der historischen Arbeit mit Urkunden.* Hg. Georg Vogeler. Köln/Wien/Weimar: Böhlau, 2009. 255–270.
- Cartelli, Antonio und Marco Palma. »Digistylus. An Online Information System for Palaeography Teaching and Research.« *KPDZ 1.* 123–134.
- Deutsch-französische Materialien für den Geschichts- und Geographieunterricht (DeuFraMa).* Braunschweig: Georg-Eckert-Institut für internationale Schulbuchforschung, 2003–2010. <<http://www.deuframmat.de/>>.
- Digitale Diplomatie. Neue Technologien in der historischen Arbeit mit Urkunden.* Hg. Georg Vogeler. Köln/Weimar/Wien: Böhlau, 2009.
- Dubs, Rolf. »Konstruktivismus: Einige Überlegungen aus der Sicht der Unterrichtsgestaltung.« *Zeitschrift für Pädagogik* 41 (1995): 889–903.
- Flindt, Nicole. *E-Learning. Theoriekonzepte und Praxiswirklichkeit.* Diss. Universität Heidelberg, 2005. <<http://www.ub.uni-heidelberg.de/archiv/6907>>.
- Geldsetzer, Sabine und Meret Strothmann. »Blende(n)d Lernen in Bochum. Integration von E-Learning in den BA/MA-Studiengang Geschichte an der Ruhr-Universität Bochum.« *Geschichte lehren an der Hochschule. Reformansätze, Methoden, Praxisbeispiele.* Hg. Rainer Pöppinghege. Schwalbach: Wochenschau-Verlag, 2007. 181–193.
- Gutbrod, Martin Andreas. *Nachhaltiges E-Learning durch sekundäre Dienste.* Diss. Universität Braunschweig, 2007. <http://rzbl04.biblio.etc.tu-bs.de:8080/docportal/receive/DocPortal_document_00021377>.
- Haug, Simone und Joachim Wedekind. »Adresse nicht gefunden. Auf den digitalen Spuren der E-Teaching-Förderprojekte.« *E-Learning. Eine Zwischenbilanz – Kritischer Rückblick als*

- Basis eines Aufbruchs*. Hg. Ullrich Dittler et al. Münster: Waxmann, 2009. 19–38.
<http://www.waxmann.com/fileadmin/media/zusatztexte/2172Volltext.pdf>.
- Henri, France et al. »E-Science, E-Research and E-Learning. New Perspectives for Graduate Studies.“ *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. Hg. Theo Bastiaens und Saul Carliner. Chesapeake: AACE, 2007. 944–949. Online: <http://www.editlib.org/p/26455>.
- Iske, Stefan und Winfried Marotzki. »Wikis. Reflexivität, Prozessualität und Partizipation.« *Medienbildung in neuen Kulturräumen. Die deutschsprachige und britische Diskussion*. Hg. Ben Bachmair. Wiesbaden: Verlag für Sozialwissenschaften, 2010. 141–151.
- Kamp, Silke. »Handschriften lesen lernen.« *KPDZ 1*. 111–122.
- Kerres, Michael. »Potenziale von Web 2.0 nutzen.« *Handbuch E-Learning. Expertenwissen aus Wissenschaft und Praxis*. Hg. Andreas Hohenstein und Karl Wilbers. München: DWD, 2006.
<http://mediendidaktik.uni-duisburg-essen.de/system/files/web20-a.pdf>.
- KPDZ 1: *Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Hg. Malte Rehbein, Patrick Sahle und Torsten Schaßan. Norderstedt: Books on Demand, 2009. Online: <http://kups.ub.uni-koeln.de/volltexte/2009/2939/>.
- Kränzle, Andreas und Gerold Ritter. *Ad fontes. Zu Konzept, Realisierung und Nutzung eines E-Learning-Angebots*. Diss. Universität Zürich, 2004. <http://www.dissertationen.uzh.ch/>.
- Latin Palaeography*. The National Archives. Kew, Richmond, Surrey: The National Archives <http://www.nationalarchives.gov.uk/latinpalaeography>.
- Mandl, Heinz und Ulrike-Marie Krause. »Lernkompetenz für die Wissensgesellschaft.« *Forschungsbericht Nr. 145*. München: Ludwig-Maximilians-Universität, Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie, 2001.
http://epub.ub.uni-muenchen.de/253/1/FB_145.pdf.
- Mankel, Mirco. *Lernstrategien und E-Learning. Eine empirische Untersuchung*. Hamburg: Kovač, 2008.
- Medieval Unicode Font Initiative*. Odd Einar Haugen et. al. [Bergen u.a.], 2001–2010.
<http://www.mufi.info/>.
- Nichols, Stephen G. »Why Material Philology? Some Thoughts.« *Zeitschrift für Deutsche Philologie* 116. Sonderheft: *Philologie als Textwissenschaft. Alte und Neue Horizonte*. Hg. Helmut Tervoren und Horst Wenzel. Berlin: Erich-Schmidt-Verlag, 1997. 10–30.
- Nikolopoulos, Alexander Stergios. *Die Sicherung der Nachhaltigkeit von E-Learning-Angeboten in Hochschulen*. Diss. Universität Frankfurt a. M., 2009.
<http://publikationen.ub.uni-frankfurt.de/volltexte/2009/7258/>.
- Paläographie Online. Von der römischen Antike bis zum Ende des Handschriftenzeitalters (1.–16. Jahrhundert)*. München: Ludwig-Maximilians-Universität, Historisches Seminar, Abt. Geschichtliche Hilfswissenschaften / Erlangen: Friedrich-Alexander-Universität Erlangen, Professur für Lateinische Philologie des Mittelalters und der Neuzeit – Virtuelle Hochschule Bayern: 2003–2010. <http://www.palaeographie-online.de/>.
- Paläographisches Lesetraining für lateinische Schriften des 5.–20. Jahrhunderts*. Thomas Frenz. Passau: Universität Passau, 2001–2005.
<http://www.phil.uni-passau.de/histhw/palaeographie>.

- Podcasts des E-Learning-Centers der Universität Zürich. E-Learning und E-Research. Zürich: Universität Zürich, 2008.
 <<http://blogs.uzh.ch/elearningpodcast/category/e-learning-und-e-research>>.
- Matthias Rohs: »Topic 05: E-Learning und E-Research.« *Podcasts des E-Learning-Centers der Universität Zürich. E-Learning und E-Research*. Zürich: Universität Zürich, 2008
 <<http://blogs.uzh.ch/elearningpodcast/2008/08/28/topic-05-e-learning-und-e-research>>.
- Reinmann, Gabi. *Selbstorganisation im Netz. Anstoß zum Hinterfragen impliziter Annahmen und Prämissen* (Arbeitsberichte Nr. 18). Augsburg: Institut für Medien- und Bildungstechnologie der Universität Augsburg, 2008.
 <http://www.imb-uni-augsburg.de/files/Arbeitsbericht_18.pdf>.
- Ruf, Urs, Nicole Frei und Tobias Zimmermann. »Leitfaden für den ICT-Einsatz in kooperativen und dialogischen Lehr-Lern-Umgebungen.« *Beiträge zur Lehrerbildung* 21/2 (2003): 192–205.
- Schmale, Wolfgang et al. *E-Learning Geschichte*. Wien/Köln/Weimar: Böhlau, 2007.
- Schulmeister, Rolf. »Studierende, Internet, E-Learning und Web 2.0.« *E-Learning 2009. Lernen im Digitalen Zeitalter*. Hg. Nicolas Apostolopoulos et al. Münster: Waxmann, 2009. 129–140.
 <<http://www.waxmann.com/fileadmin/media/zusatztexte/2199Volltext.pdf>>
- Seufert, Sabine und Dieter Euler. *Nachhaltigkeit von eLearning-Innovationen. Fallstudien zu Implementierungsstrategien von eLearning als Innovationen an Hochschulen*. SCIL-Arbeitsbericht 4). St. Gallen: SCIL, 2005. <<http://www.scil.ch/fileadmin/Container/Leistungen/Veroeffentlichungen/2005-01-seufert-euler-nachhaltigkeit-elearning.pdf>>.
- Thaller, Manfred. *Grenzen der Digitalisierung?* [Podiumsgespräch vom 5. Februar 2010 an den Schweizer Geschichtstagen 2010.] Basel: Soundcloud, 2010.
 <<http://soundcloud.com/infoclio-ch/podium-grenzen-der-digitalisierung>>.
- Teuscher, Simon. »Bernische Privatbriefe aus der Zeit um 1500. Überlegungen zu ihren zeitgenössischen Funktionen und zu Möglichkeiten ihrer Auswertung.« *Mittelalterliche Literatur im Lebenszusammenhang. Ergebnisse des Troisième Cycle Romand 1994*. Hg. Eckart Conrad Lutz. Freiburg: Universitätsverlag Freiburg, 1997. 359–385.
- Teuscher, Simon. »Threats from Above on Request from Below. Dynamics of the Territorial Administration of Berne, 1420–1450.« *Empowering Interactions. Political Cultures and the Emergence of the State in Europe, 1300–1900*. Hg. Wim Blockmans, André Holenstein und Jon Mathieu. Farnham: Ashgate, 2009. 101–114.
- Zimmermann, Tobias et al. »Dialog mit 200 Studierenden – geht das? Blended Learning in einer Vorlesung mit hoher Teilnehmerzahl.« *Das Hochschulwesen* 6 (2008): 179–185.

L'édition électronique de cahiers de travail : l'exemple de Mes Pensées de Montesquieu

Carole Dornier, Pierre-Yves Buard

Résumé

S'appuyant sur l'exemple du projet Montedite, une édition électronique d'un cahier de travail du célèbre auteur de *L'Esprit des lois*, ce chapitre vise à montrer que baliser des manuscrits et prévoir des liens entre les images numériques et la transcription, page par page, change la façon dont nous considérons et classifions ce manuscrit. Cette méthode offre, dans ce cas particulier, le moyen de fournir, avec une transcription fidèle, une analyse chronologique de la pensée et de la documentation de Montesquieu, grâce à une façon aisée, à côté des parties autographes, d'identifier des mains différentes et datées (les secrétaires engagés par l'auteur). Cette édition électronique utilisant les balises TEI pour sources primaires montre ce cahier de travail qui accompagne l'auteur pendant trente ans comme une réserve utilisée pour conserver des énoncés, des informations et des idées. Des renvois dynamiques et des index soulignent la fonction exacte de ce manuscrit. Les facsimilés des manuscrits permettent d'étudier ces pratiques qui ont joué un rôle important dans la création intellectuelle et culturelle : *excerpta*, recueils de lieux communs, cahiers de travail et carnets de notes, et différentes sortes de compilation personnelle. La numérisation et l'édition numérique ne sont pas seulement des outils pour la critique génétique à propos du processus d'écriture des auteurs. Elles permettent d'explorer et d'analyser pour les chercheurs et pour un public plus large comment des hommes du passé, avec du papier et de l'encre, créaient des instruments élaborés pour conserver, classer et utiliser les savoirs, sans ordinateurs.

Zusammenfassung

Anhand des Montedite Projekts, der digitalen Edition eines Arbeitsheftes des französischen Staatstheoretikers Montesquieu, soll dieser Artikel beispielhaft aufzeigen, wie die Kodierung seiner handschriftlichen Notizen und die seitengetreue Verknüpfung digitaler Faksimiles und Transkriptionen die Wahrnehmung und Bewertung seines Werkes verändert und eine chronologische Analyse von Montesquieus Denken und Schaffen ermöglicht. Neben den Notizen des Autors selbst lassen sich Vermerke seiner zu datierbaren Zeitpunkten wechselnden Sekretäre identifizieren. Basierend auf einer TEI kodierten Transkription macht die Edition ein Arbeitsheft zugänglich, das seinem Besitzer dreizehn Jahre lang dazu diente, Ideen und Informationen in Form von Notizen

festzuhalten. Als Link realisierte Querverweise und Indizes erfüllen die ursprüngliche Funktion der Handschrift. Anhand der Faksimiles läßt sich diese als kulturelle und intellektuelle Schaffenspraxis nachvollziehen: das handschriftliche Exzerpieren und Kompilieren. Digitale Reproduktion und Edition des Arbeitsheftes dienen nicht allein der Analyse des Entstehungsprozesses von Montesquieus Werk. Sie gewähren Forschern und Öffentlichkeit eine Einsicht darin, dass Papier und Tinte zweckdienliche Werkzeuge für die Verarbeitung von Wissen und Information sein konnten, bevor sie durch den Computer weitestgehend verdrängt wurden.

Abstract

Leaning on the example of the Montedite project, an electronic edition of a notebook by the famous author of *L'Esprit des lois*, this chapter aims at showing that encoding manuscripts and providing links between digital images and transcriptions, page by page, changes the way we consider and classify this manuscript. This method offers, in this particular case, the means to provide, together with an accurate transcription, a chronological analysis of Montesquieu's thought and documentation. It allows the reader, to identify different handwritings and dates. Based on a TEI XML-encoded transcription, this web edition makes accessible a notebook which accompanied the author's career for thirteen years to keep notes, information and ideas. Cross-references and indexes in the edition carry out the genuine function of the manuscript. Facsimiles of the manuscripts give an opportunity to study these practices which have played an important part in cultural and intellectual creation: *excerpta*, common-place books, workbooks and notebooks, and different kinds of personal compilations. Digitisation and digital editions are not only tools for genetic criticism about authors' writing processes. They allow researchers and a larger public to explore and analyze how men in the past with paper and ink created elaborate instruments to store, classify and use knowledge without computers.

1. Introduction

L'accès au patrimoine manuscrit par la numérisation des originaux offre aux chercheurs l'occasion de modifier la façon dont ils conçoivent leur travail et leur rôle dans l'espace public et social. Jusqu'à aujourd'hui l'éditeur scientifique était un médiateur entre des originaux difficilement accessibles et un public pour lequel cette médiation était nécessaire. Les savants étaient le lien entre des documents silencieux dissimulés dans les fonds patrimoniaux des bibliothèques ou dans les archives et le savoir plus ou moins diffusé dans le public par l'édition d'un contenu signifiant : le texte. L'arrivée sur internet des images numérisées a profondément changé cette relation. Si quelques manuscrits

seulement sont jusqu'à maintenant disponibles en ligne, la question se pose de profiter de cette stimulante possibilité pour montrer l'objet matériel à côté de la transcription comme texte lisible.

2. La spécificité du manuscrit des *Pensées*

Cette opportunité semblerait favoriser surtout les recherches sur les manuscrits de la période contemporaine. En effet les études de génétique portant sur des manuscrits des XIX^e, XX^e et XXI^e siècles sont stimulées par l'abondance des matériaux à la disposition du chercheur. Avec la sacralisation de la figure de l'auteur, l'intérêt pour toutes les traces écrites du processus de l'œuvre a contribué à leur conservation. La critique génétique s'est intéressée très tôt aux usages de l'informatique dans l'exploitation des dossiers d'une œuvre. Aujourd'hui, pour ne parler que du domaine français qui a donné naissance aux travaux pionniers de l'ITEM, les manuscrits de Beckett, ceux de Stendhal et de Flaubert, de Proust, de Roland Barthes ont donné lieu à divers projets d'éditions électroniques et de bases de données qui exploitent la confrontation texte/image grâce au numérique. Cette importance des matériaux manuscrits à disposition et les possibilités offertes par ce biais aux études génétiques ont favorisé les projets portant sur des œuvres littéraires des XIX^e et XX^e siècles et sur tout ce qui permet d'éclairer un processus de création originale.

Pour les époques précédentes, on trouve très peu de manuscrits d'auteur, la publication d'un ouvrage entraînant, à de très rares exceptions près, la destruction du manuscrit. Le manuscrit de Montesquieu intitulé *Mes Pensées* est d'autant plus intéressant. Il s'agit d'un cahier de travail en trois volumes consulté et enrichi pendant trente ans, entre 1726–1727 et 1754, par le célèbre auteur de *L'Esprit des Loix*. Il comporte environ 1100 feuillets. N'étant pas destiné à la publication, il n'existe que comme manuscrit. Instrument de documentation personnelle et de composition, il peut être considéré comme le « laboratoire de l'œuvre » d'un auteur majeur des Lumières. Une autre de ses caractéristiques est de comporter une importante partie autographe et différents scripteurs, les secrétaires de Montesquieu, quelques interventions très brèves et rares de Jean-Baptiste de Secondat, le fils de l'auteur et des écritures plus tardives dans les marges, révélant la trace de lectures posthumes à des fins d'éditions partielles ou intégrales (fig. 1).

Par ailleurs, conçu comme espace de conservation de fragments utilisés dans des œuvres imprimées ou laissés en attente d'utilisation, le manuscrit signale en marge les passages utilisés dans les œuvres publiées ou projetées de l'auteur. Ainsi l'expression récurrente en marge « Mis dans les Loix » par exemple, signifie qu'un élément de texte a été utilisé dans un chapitre de *L'Esprit des lois* (fig. 2). En outre le manuscrit comporte

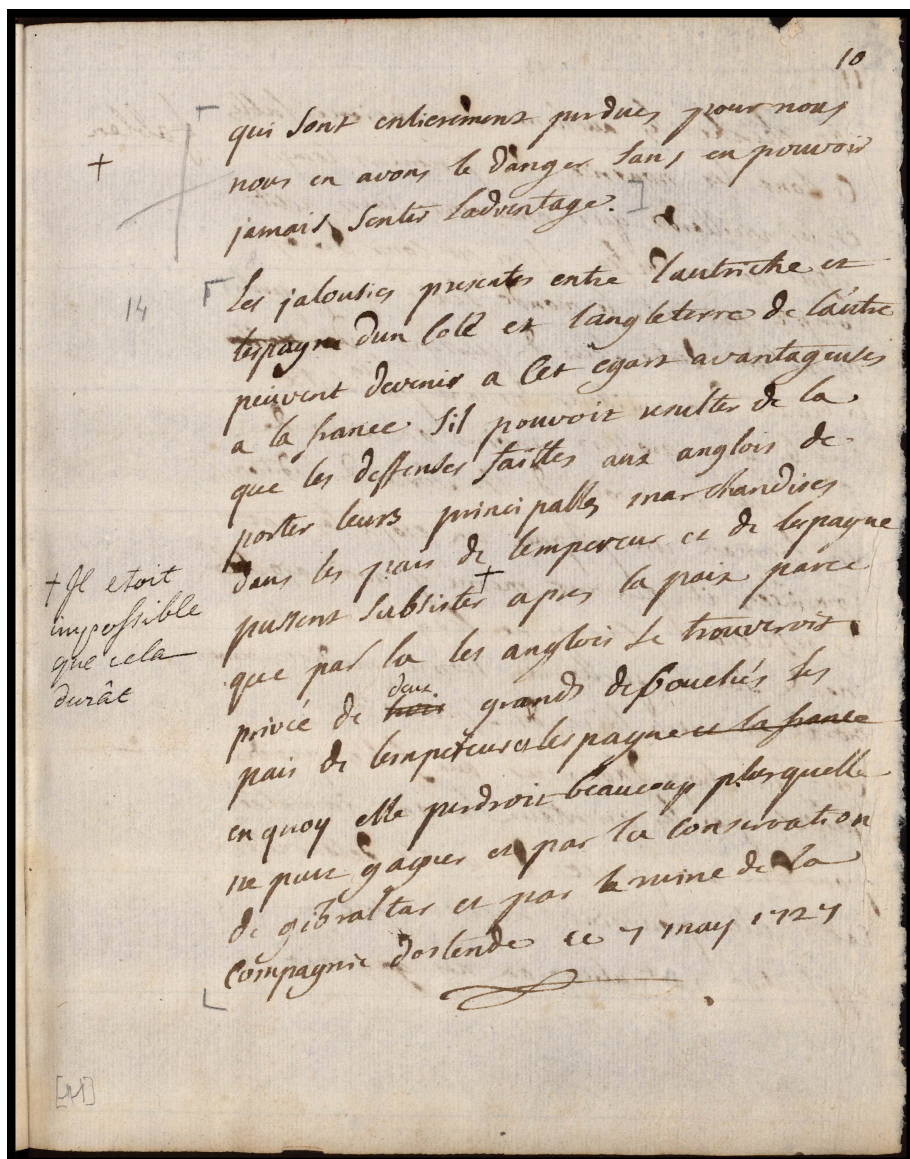


FIGURE 1. Montesquieu, *Mes Pensées*, BM Bordeaux, ms. 1866, t. 1, p. 10. Ensemble autographe (1727) avec note postérieure en marge gauche écrite par un secrétaire (1743–1744).

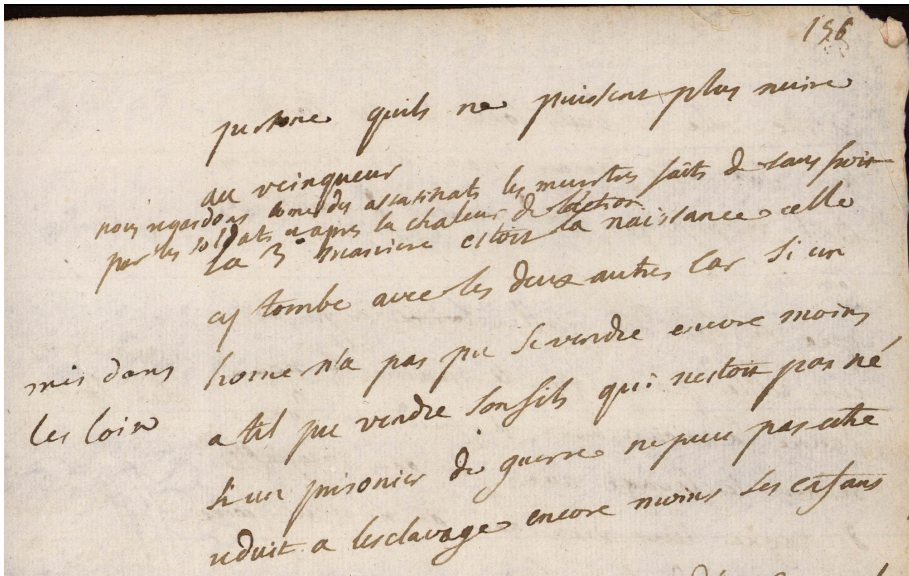


FIGURE 2. Montesquieu, *Mes Pensées*, B. M. Bordeaux, ms. 1866, t. 1, p. 156 (détail). En marge gauche : « Mis dans les loix ».

de fréquents renvois internes qui permettent de repérer le traitement d'une même thématique à des endroits différents de cet ensemble.

Les caractéristiques du manuscrit rendent la consultation de l'original particulièrement nécessaire à la compréhension de tout ce que ce recueil peut nous apprendre sur la genèse des œuvres publiées, sur les projets et les ébauches non abouties qu'il est possible d'inventorier et de dater. L'identification des écritures est un moyen de préciser les étapes de l'information, de la culture, de la pensée de Montesquieu. En effet l'écriture de chaque secrétaire, au service de l'écrivain pendant une période déterminée que des chercheurs ont définie en particulier grâce à la correspondance, datée, donne des indications sur les années au cours desquelles tel ou tel fragment a été transcrit. L'examen attentif et renouvelé du manuscrit est donc le point de départ indispensable d'une réflexion rigoureuse, en rupture avec une utilisation abusive de fragments non datés.

3. Le choix d'une confrontation texte/image et d'un balisage TEI

La numérisation permet au chercheur d'examiner à l'infini cet original pour confirmer des hypothèses dans la chronologie de la composition et de la documentation de l'auteur.

Elle constitue une aide à la transcription grâce à des images haute définition. Mais l'image numérique permet aussi d'envisager un autre mode d'édition qui restitue les spécificités de ce cahier de travail et met en évidence une méthode intellectuelle et les étapes d'une réflexion. Le projet *Montedite* qui est devenu une édition dans le cadre d'une collection numérique des Presses universitaires de Caen, a été conçu pour fournir au lecteur, avec une transcription fidèle, une analyse chronologique de la pensée et de la documentation de Montesquieu, en particulier en identifiant, à côté des parties autographes, des mains différentes et datées (les secrétaires engagés par l'auteur). Il s'agissait de permettre la confrontation, à l'écran, de l'image numérisée et de sa transcription enrichie d'informations sur la chronologie des écritures, sur la progression de la composition par le signalement des suppressions et des additions. Cette édition en ligne met donc en relation le texte et l'image ; elle est dynamique dans la mise à disposition à l'écran des informations nécessaires à l'interprétation ; elle favorise des explorations systématisées du document. L'enrichissement de la transcription a été conçu en utilisant le XML et un balisage TEI (Text Encoding Initiative), d'abord selon la version P4, utilisée au début du projet, puis plus récemment en migrant dans la version P5, version que nous présentons ici. L'encodage a été conçu en recherchant la simplicité, l'économie de moyens et la meilleure adaptation aux spécificités de ce manuscrit qui est un état unique corrigé, avec suppressions et additions, ce qui ne correspond pas tout à fait aux dossiers plus complexes traités dans le cadre de la génétique textuelle.

Les principaux éléments utilisés sont : `<div>`, `<pb>`, `<p>`, `<add>`, ``, `<note>`, `<handNotes>`, `<handNote>`, `<handShift>`, `<title>`, `<ref>`. Les divisions originales du support matériel (volume, page ou folio, paragraphe) sont encodées pour la mise en relation texte/image, ainsi que la numérotation des fragments par le premier éditeur du texte intégral, Henri Barckhausen, numérotation que l'on utilise par commodité. L'élément `<div>` est alors utilisé avec un attribut `« xml:id »` dont la valeur est définie à partir du numéro conventionnel : `<div xml:id="pn125">`. Le balisage des corrections est conçu comme une aide à la lecture du manuscrit en mode image grâce au repérage des parties biffées (``) et des parties ajoutées (`<add>`), et à la possibilité, dans la version diffusée sur le web, de masquer les portions de texte supprimées et de lire ainsi la version finale voulue par l'auteur. L'identification des mains est encodée par les attributs `« xml:id »` et scribe des éléments `<handNotes>`, `<handNote>`. Cet encodage permet de dater un élément de texte parfois très postérieur au fragment dans lequel il s'insère. On utilise alors l'attribut `« hand »` de l'élément `<add>`. Un changement de main est signalé par l'élément `<handShift>` complété par les attributs `« new »` et `« next »`. Le statut textuel incertain des notes marginales qu'on peut considérer tantôt comme notes, tantôt comme additions, tantôt comme directives ou notes de régie, a conduit à la généralisation de la localisation de ces éléments de texte en marge droite de la transcription. Les éléments utilisés sont `<note>` et `<add>` auquel on associe l'attribut `« place »` (`<add place="en marge">`). A l'écran, l'ensemble de ces éléments peut être mis

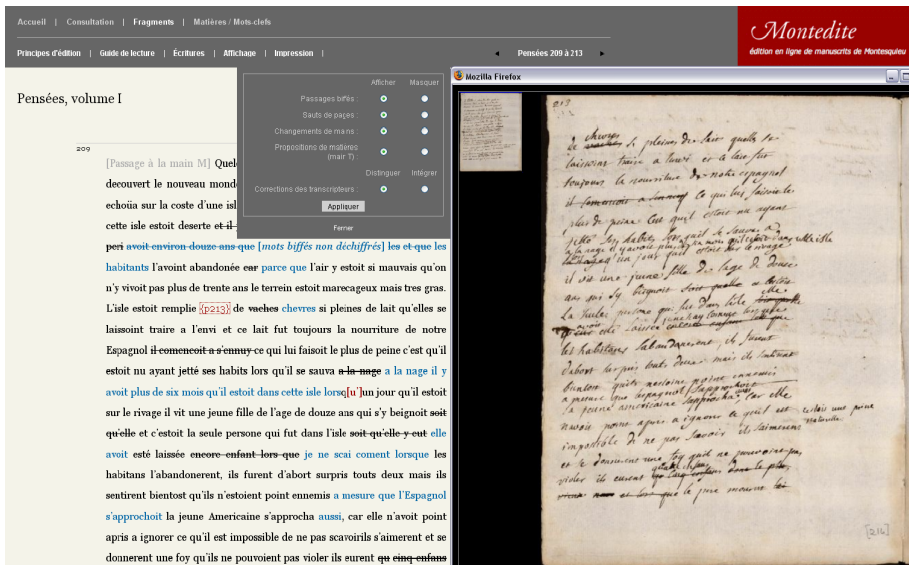


FIGURE 3. Transcription à gauche du n° 209 (*Pensées*, volume I, p. 212–213); fenêtre de la page 213 du manuscrit.

en forme dynamiquement pour faciliter la lecture ; les additions sont signalées dans la transcription par des caractères de couleur bleue. Cette édition a été conçue comme une aide à la lecture et à l'exploitation des images numérisées du manuscrit.

Un onglet « Écritures » donne accès à une autre fenêtre précisant la période d'intervention de chaque scripteur, identifié par une lettre majuscule (fig. 3). L'écriture principale du fragment à l'écran est signalée sur la première ligne en marge droite. Un simple clic sur un texte ajouté (en bleu dans l'interface web) permet de faire apparaître une pop-up qui précise l'identité du scripteur et la position de l'addition par rapport à l'écriture principale (fig. 5 : « E, en bas de page, sous le paraphe de séparation »).

4. Les résultats scientifiques attendus

La confrontation de l'image et de la transcription enrichie par balisage TEI conduit donc à une analyse du manuscrit éclairant les étapes de l'information, de la culture, de la pensée de Montesquieu, et aidant à l'interprétation de fragments par leur datation approximative.

Par exemple en consultant le n° 103 (numérotation Barckhausen), on constate sur l'image de la page correspondante du manuscrit qu'une remarque par une écriture différente de celle qui figure sur les pages 96 et 97 a été ajoutée sur un espace libre en bas de page (fig. 4).

La transcription encodée permet de dater l'addition (fig. 5).

L'addition est écrite par le secrétaire E (1734–1739), qui intervient après le voyage en Italie de Montesquieu (1728–1731), tandis que l'écriture des n° 102 et 104 est celle de Bottereau-Duval, qui n'est plus au service de l'auteur après 1731. La remarque n° 103 est donc postérieure aux voyages, insérée dans une séquence rédigée antérieurement. La comparaison du n° 103 avec son environnement textuel permettra ensuite de conjecturer les raisons pour lesquelles Montesquieu a choisi cet emplacement dans le volume 1 : le passage qui précède (n° 102) traite en effet du changement continu qui affecte le monde physique ; ce changement permet de rendre compte de ce qui apparaît comme les exagérations des anciens dans leurs évocations du monde qui les entourait. Ainsi le lieu fameux d'une célèbre victoire romaine, le lac Régille, apparaît au voyageur du XVIII^e siècle « pas plus grand que la main ». Cette pensée s'éclaire par celle qui la précède mais l'étude des écritures et de la localisation de l'addition sur la page nous renseigne aussi sur la façon de travailler de Montesquieu. Bien qu'il refuse, pour son cahier de travail, un classement systématique, il n'insère pas une remarque additionnelle tout à fait aléatoirement sur des emplacements restés libres. Il procède par rapprochements, association d'idées et d'informations.

On peut également constater qu'il revient sur certains passages pour les actualiser quand le contexte politique par exemple a changé.

Sur la page 161 du manuscrit (volume I), on remarque des corrections d'une main différente de la main principale (fig. 6). La transcription balisée permet de dater et d'expliquer ces corrections : en 1736, la Sicile qui appartenait à l'empereur d'Autriche sera attribuée à Don Carlos, fils de Philippe V d'Espagne, roi de Naples. Après 1736, Montesquieu fait remplacer par le secrétaire E toutes les mentions de l'empereur par celles du « roy de Naples », Don Carlos, nouveau possesseur de la région et, comme on le voit ici, les « troupes impériales » sont corrigées en « troupes royales ».

D'autres possibilités d'exploration du recueil ont été prévues : le manuscrit remplit en effet une fonction de stockage de fragments qui ont été utilisés dans des œuvres imprimées ou bien qui en ont été rejetés mais que l'auteur a souhaité conserver. L'encodage de ces mentions d'emplois ou de rejets permet de systématiser les références à des œuvres publiées ou projetées comme celle-ci : « J'ai mis cela dans mes Lois » [*i.e.* « J'ai utilisé ce passage dans *L'Esprit des lois* »]. On utilise l'élément <note> complété par l'attribut type pour marquer de telles mentions marginales, ce qui donne pour cet exemple : <note type="RU_EL"> (EL, abréviation pour *L'Esprit des lois*). L'intérêt de cet encodage est de pouvoir extraire de façon systématique toutes les mentions de ce type, toutes celles qui renvoient à la même œuvre, de pouvoir aussi établir la liste de toutes les

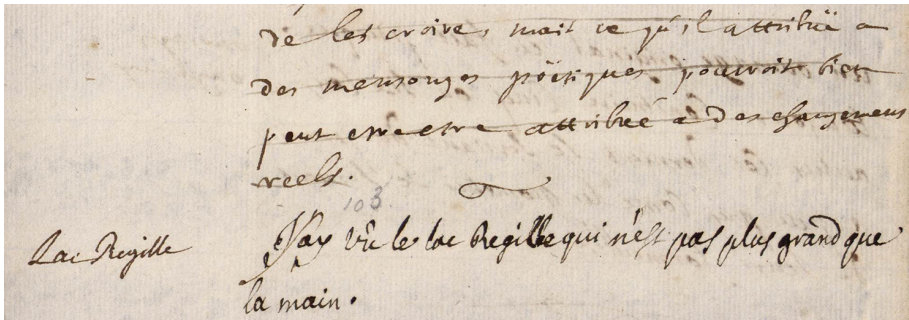


FIGURE 4. N° 103, page 96 du volume I des *Pensées* (détail). En bas : « J'ay vu le lac Regille qui n'est pas plus grand que la main ».

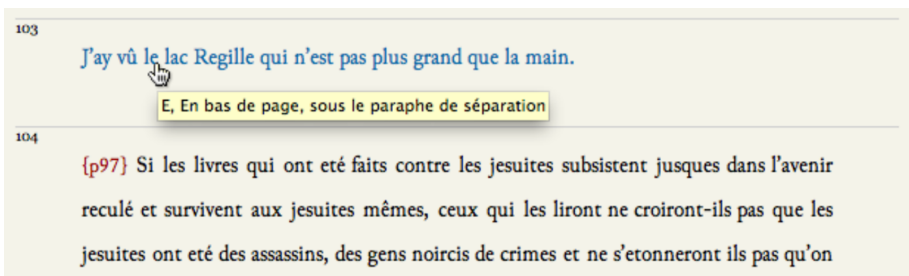


FIGURE 5. Transcription du n° 103. La couleur bleue signale une addition postérieure à la période de transcription de la séquence. La pop-up identifie le scripteur et permet la datation.

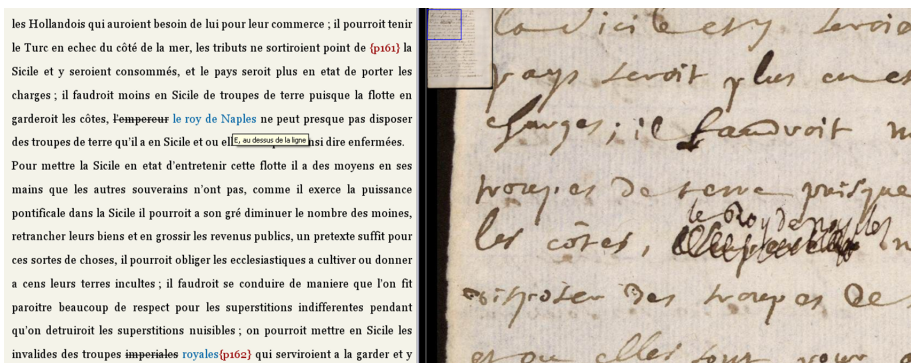


FIGURE 6. Transcription du n° 177 et fenêtre de la page du manuscrit (volume I, p. 161). La pop-up ouverte en cliquant sur « le roy de Naples » (en bleu dans le texte) indique : « E, au dessus de la ligne ».

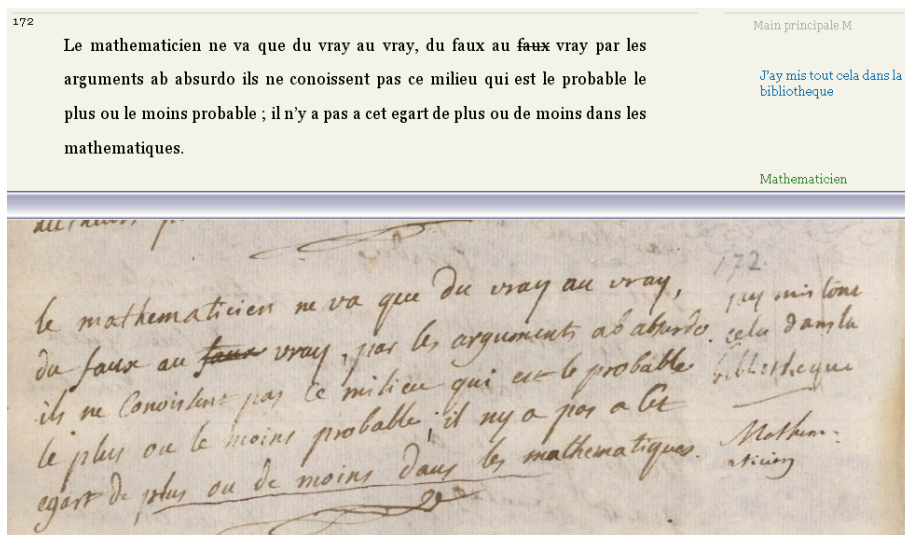


FIGURE 7. En marge droite (n° 172) : « J'ay mis tout cela dans la bibliotheque ».

œuvres de l'auteur mentionnées, de se doter des moyens d'un accès automatique à ces mentions par nom d'œuvre. L'encodage permet aussi de signaler les traces des œuvres abandonnées ou les références aux lectures de l'auteur. Le balisage a pour fonction de systématiser l'extraction pour constitution de listes : des ouvrages allographes sources (<bibl>), des extraits de lecture mentionnés, des projets d'ouvrages qui se réduisent parfois à un titre (<title>). Il permet en outre d'activer les renvois internes prévus par l'auteur (<ref> : définit une référence vers un autre emplacement). Le balisage de la transcription du manuscrit permet non seulement d'éclairer la genèse des œuvres publiées en repérant les passages transcrits au préalable dans les *Pensées*. Il permet aussi d'inventorier en les datant les projets et les ébauches non aboutis, leur transformation dans des écrits orientés différemment. Le n° 172 (volume I) offre un exemple de ce moyen de construire une étude de genèse par analyse du manuscrit (fig. 7).

La note marginale fait état d'une réutilisation du fragment dans un ouvrage intitulé *Bibliothèque*. La confrontation rapide des titres contenus dans le manuscrit grâce au balisage permet de faire apparaître un projet de recueil de notes ou d'ouvrage à contenu historique et politique désigné par plusieurs titres : *Princes* (n° 540, 610) ou *Prince* (n° 640), *Bibliothèque espagnole* (n° 524–526), *Bibliothèque* (n° 173), *Journal* (n° 140, 162, 194, 318, 478), *Journal espagnol* (n° 472), à rapprocher du manuscrit *Réflexions sur le*

[Accueil](#) | [Consultation](#) | [Fragments](#) | [Matières / Mots-clefs](#)

Montedite
édition en ligne de manuscrits de Montesquieu

[Principes d'édition](#) | [Guide de lecture](#) |

Résultats de la recherche

Recherche de **Espagnol** dans **toute la base**

Il y a **5** occurrences correspondantes à votre recherche

1. *Pensées, volume I* [[Pensée 38](#)]
n'étoit pas fait comme du terme d'Hercule. ----- 33 Les peuples de ce continent de l'Amérique qui est entre le pays espagnol et anglois nous donnent l'idée de ce qu'étoient les premiers hommes avant l'établissement des grandes sociétés et la culture des terres. Les peuples chasseurs sont ordinairement anthropophages ils

2. *Pensées, volume I* [[Pensée 170](#)]
----- Remarqués que le bon foy des Espagnols a ruiné leur commerce et l'a transporté aux étrangers qui le font sans aucune crainte sous le nom d'un Espagnol. ----- (J'aime les querelles sur les ouvrages des anciens et des modernes elles prouvent qu'il y a d'excellens auteurs parmi les

3. *Pensées, volume I* [[Pensée 209](#)]
de vaches chevrès si pleines de lait qu'elles se laissent traire à l'envi et ce lait fut toujours la nourriture de notre Espagnol il commençoit à s'ennuyer ce qui lui faisoit le plus de peine c'est qu'il estoit nu ayant jeté ses habits lors qu'il se sauva à la nage il y avoit plus de six mois

4. *Pensées, volume I* [[Pensée 209](#)]

FIGURE 8. Résultats d'une recherche à partir du mot *espagnol*.

caractère de quelques princes et sur quelques événements de leur vie (ca. 1731–1733). Cet ensemble nous renseigne sur la façon de travailler de Montesquieu, qui ouvre, sur des sujets variés, des dossiers aux contours changeants, qui n'aboutissent pas toujours à des écrits autonomes et achevés et dont des éléments sont réutilisés dans des pièces diverses.

5. Les instruments de navigation à destination de publics élargis

Le site Montedite offre aux chercheurs, étudiants et public un moyen de naviguer dans le texte des *Pensées* et de faire des recherches plein texte. Le moteur de recherche PhiloLogic™ offre l'opportunité d'une recherche de mots avec variations orthographiques, particulièrement précieuse pour un texte présentant des différences orthographiques selon les scripteurs. On trouve ci-dessous (fig. 8) l'exemple d'une recherche à partir du mot *espagnol*.

Comme on le voit ici, la recherche par terme fournit des résultats avec l'environnement textuel. Pour chaque occurrence, le numéro (ex : *Pensée 472*), par un lien, permet d'accéder au passage du texte correspondant.

Enfin la reprise des thèmes notés en marge par un lecteur du XIX^e siècle chargé par la famille de Montesquieu d'une édition qui n'a jamais abouti a permis de constituer un index thématique facilitant les recherches dans les volumes (fig. 9).

Accueil
Consultation
Fragments
Matières / Mots-clefs

Principes d'édition
Guide de lecture

Montedite
édition en ligne de manuscrits de Montesquieu

M

M^r de Lamothe : 116
Mad^e Dacier : 116
Mahomet : 503
Mahométans Mariage : 83
Mal vénérien : 216
Maladies : 86
Malebranche : 156, 410, 436
Malheur fait chercher Dieu : 390
Malheureux : 457
Marbres : 399

Monarchie : 597
Monnoie a fait les grands empires : 647
Montagne : 633
Montres : 529
Mort : 349
Mort de Neron dans Suetone : 698
Mort des princes : 432
Mort d'Alexandre : 99
Mort et genre de mort : 641
Mort pour un Romain : 646

FIGURE 9. Index thématique reprenant les mots-clefs figurant dans les marges du manuscrit.

Les numéros qui suivent les mots-clefs (numérotation Barckhausen) permettent d'accéder au texte par un lien.

L'accès libre en accord avec la bibliothèque de Bordeaux, détentrice du manuscrit, est un élément essentiel de valorisation de cette pièce inestimable, valorisation qui peut intéresser chercheurs, étudiants et un public plus large, curieux de ce qui concerne Montesquieu mais aussi des manuscrits d'auteurs de l'âge classique.

6. Conclusion

Ce manuscrit particulier, qui relève en partie d'un héritage humaniste de compilation et de documentation et de la tradition des recueils de lieux communs (Dornier 2008 : 809–20), souligne l'intérêt de l'édition électronique pour étudier ces pratiques qui ont joué un rôle important dans la création intellectuelle et culturelle : *excerpta*, recueils de lieux communs, cahiers de travail et carnets de notes, et différentes sortes de compilation personnelle. La numérisation et l'édition numérique, comme cela est souligné par différents projets concernant des encyclopédies médiévales, des carnets de notes et des écrits scientifiques (*Newton Project*, *Sourcencyme*, *Cybernard* etc...) ne sont pas seulement des outils très intéressants pour la critique génétique à propos du processus d'écriture littéraire des écrivains. Elles permettent d'explorer et d'analyser pour les chercheurs et pour un public plus large comment des hommes du passé, avec du papier

et de l'encre, créaient des instruments élaborés pour conserver, classer et utiliser les savoirs, sans ordinateurs.

Bibliographie

- Benrekassa, Georges. « Les Manuscrits de Montesquieu, Secrétaires, Écritures, Datations ». *Cahiers Montesquieu* 8 (2004).
- Cybernard – ANR. Institut des Textes et Manuscrits Modernes (ITEM). Dir. Jean-Louis Lebrave. CNRS/ENS – UMR 8132. <<http://www.item.ens.fr/index.php?id=14060>>.
- De Biasi, Pierre-Marc. *La Génétique des textes*. Paris : Armand Colin, 2005.
- Décultot, Elisabeth. *Lire, copier, écrire. Les bibliothèques manuscrites et leurs usages au XVIIIème siècle*. Paris : CNRS Editions, 2003.
- Dornier, Carole. « La Mise en archive de la réflexion dans les *Pensées* ». *Revue Montesquieu* 7 (2003–2004) : 25–39. En ligne : <<http://montesquieu.ens-lyon.fr/spip.php?article413>>.
- Dornier, Carole. « Montesquieu et la tradition des recueils de lieux communs ». *Revue d'Histoire littéraire de la France* 4 (2008) : 809–820.
En ligne : <<http://hal.archives-ouvertes.fr/hal-00348800>>.
- Grésillon, Almuth et Jean-Louis Lebrave. *Ecrire au XVIIe et XVIIIe siècles. Genèses de textes littéraires et philosophiques*. Paris : CNRS Editions, 2000.
- ITEM : Institut des Textes et Manuscrits Modernes. Laboratoire du CNRS constitué en Unité Mixte de Recherche CNRS (Centre National de Recherche Scientifique) / ENS (École Normale Supérieure) – UMR 8132, 2009–2010. <<http://www.item.ens.fr/>>.
- Lebarbé, Thomas et Cécile Meynard. « Nouvelles pratiques éditoriales, nouvelles lectures : Les enjeux de l'édition électronique de manuscrits littéraires ». *Mémoires Du Livre*. Vol. 1.1 (2009). <<http://www.erudit.org/revue/memoires/2009/v1/n1/038635ar.html>>.
- Lerich, Françoise et Cécile Meynard (dir.). « De L'hypertexte au manuscrit. L'apport et les limites du numérique pour l'édition et la valorisation de manuscrits littéraires modernes ». *Recherches & Travaux* 72 (2008) : 9–301.
- Montedite. *Édition électronique des Pensées de Montesquieu*. Ed. Pierre-Yves Buard et Carole Dornier. Caen : Société Montesquieu, Maison de la Recherche en sciences humaines de l'Université de Caen, Presses universitaires de Caen, 2006–2011.
<<http://www.unicaen.fr/services/puc/sources/Montesquieu/>>.
- Montesquieu. « Spicilège ». *Œuvres complètes*, t. 13. Ed. Rolando Minuti et Salvatore Rotta. Oxford : The Voltaire Foundation, Napoli, Istituto Italiano per gli Studi Filosofici, 2002. 37–77.
- Newton Project. Eds. Rob Iliffe and Scott Mandelbrote. Sussex : University of Sussex, 1998–2010. <<http://www.newtonproject.sussex.ac.uk/>>.
- PhiloLogic™. Moteur de recherche développé par le ARTFL Project et le Digital Library Development Center (DLDC) à l'University de Chicago. Chicago : University of Chicago, 2010. <<http://sites.google.com/site/philologic3/>>.
- Shackleton, Robert. « Les secrétaires de Montesquieu ». *Montesquieu, Œuvres complètes*. Paris : Nagel (1950). II, xxxv–xlili.

Sourcencyme – ANR. Atelier Vincent de Beauvais. Dir. Isabelle Draelants. Nancy : Université de Nancy 2, 2007–2010.

<<http://www.univ-nancy2.fr/MOYENAGE/VincentdeBeauvais/ProgrammeSources.html>>.

TEI : Text Encoding Initiative. TEI Consortium, 2010. <<http://www.tei-c.org/index.xml>>.

Volpillac-Auger, Catherine. « De la main à la plume. Montesquieu et ses secrétaires : une mise au point ». *Montesquieu en 2005. Studies on Voltaire and the eighteenth century* 05 (2005) : 103–151.

Archives d'un lecteur philosophe. Le traitement numérique des notes de lecture de Michel Foucault

Samantha Saïdi, Jean-François Bert, Philippe Artières

Résumé

Nous détaillerons le travail effectué sur une archive de pièces manuscrites : essentiellement des fiches de lecture, des fiches thématiques et des fiches bibliographiques, que le philosophe Michel Foucault rédigea et rassembla en un instrument de travail notionnel pour la rédaction de son essai, *Les mots et les choses*. Cet échantillon de sa bibliothèque de travail privée constitue un véritable champ d'investigation de la façon dont Michel Foucault utilisait ses sources primaires et secondaires : comment il les lisait, les transcrivait et, enfin, les restituait dans ses ouvrages. Bien que ce travail d'investigation ait pour matériau principal un objet qui l'exclut a priori d'une approche codicologique (codex, parchemins vs papier du XXème s.) ou paléographique (époque), nos méthodes sont proches et nous voudrions partager ici avec les spécialistes de ces disciplines, notre approche propre à l'anthropologie des pratiques culturelles, mais surtout notre expérience de manipulation des outils et méthodes de l'archivistique, de l'informatique et du numérique. En effet, de la Description Archivistique Encodée (EAD) au Langage de Modélisation Unifiée (UML), ces outils et méthodes méritent d'être décrits car c'est grâce à eux que ce corpus qui donne à voir la « bibliothèque » choisie par le philosophe pour la rédaction des *Mots et les choses*, est aujourd'hui intelligible et surtout utilisable pour la recherche.

Zusammenfassung

Dieser Beitrag erläutert, auf welche Weise ein Team aus Anthropologen und Informationswissenschaftlern die handschriftlichen Vorarbeiten Michel Foucaults zu *Les mots et les choses* (dt.: *Die Ordnung der Dinge*) erschlossen haben: die Digitalisierung eines Konvoluts aus Exzerpten, thematischen Aufzeichnungen und bibliographischen Notizen. Es handelt sich bei dieser privaten »Arbeitsbibliothek« um eine ideale Quellenbasis, um Aufschlüsse über Foucaults Schaffensprozess und seinen Umgang mit Primär- und Sekundärquellen zu erlangen: Wie hat Foucault relevante Texte gelesen, transkribiert, exzerpiert und in seinem Werk weiterverwendet? Obgleich es sich um zeitgenössische Notizen auf Papier und somit nicht um einen Gegenstand kodikologischer oder paläographischer Forschung im engeren Sinne handelt, ist der methodische Zugriff derselbe oder

doch zumindest weitgehend übertragbar und soll aus diesem Grund in diesem Band zur Diskussion gestellt werden. Mögen die Ausgangsfragen einem anthropologischen Interesse an kulturellen Praktiken entspringen, so geht es bei der Quellenerschliessung in erster Linie um die Verwendung digitaler archivalischer Hilfsmittel und Methoden. Von der Verwendung des XML-Standards EAD (Encoded Archival Description) zur vereinheitlichten Modellierungssprache UML (Unified Modeling Language) werden die Methoden vorgestellt, dank derer Foucaults Auswahlbibliothek zur Vorbereitung von *Les mots et les choses* in einem virtuellen Archiv sowohl der Forschung als auch der interessierten Öffentlichkeit zugänglich gemacht werden konnten.

Abstract

In this chapter we explain how our team of anthropologists and information scientists worked on Michel Foucault's characteristic set of handwritten material: mainly reading cards, bibliographical and subject cards that Michel Foucault wrote, filed and used before writing his essay *Les mots et les choses* (*The Order of Things*). This digitised sample of his private "working library" constitutes an exceptional medium to investigate, discover and analyze how Foucault used primary and secondary sources for his writing: how he read them, extracted from them, adapted them in his handwritten transcriptions and in the end transformed these transcriptions into his final work. Even though this handwritten material is contemporary and thus does not convey an object for codicological or palaeographical research in a strict sense, our perspective and the methodology and techniques applied are similar. This sample was indeed an ideal corpus for testing the use of digital and data-processing methods and tools. By these means we were able to keep a digitised record of the archive, by describing, annotating and extracting indexes, but they also gave us the means of modeling in UML the final web application that will bring together the EAD based catalogue and the virtual Foucault private library. The portal will enable researchers and, to some extent, the general public to see, understand and use the "library" chosen by the philosopher for the drafting of *Les Mots et les choses*.

Introduction

Dans un texte intitulé « À propos des faiseurs d'histoire », autour d'une affaire de plagiat qui touche un livre de Jacques Attali, Michel Foucault indique :

[C]e qui me paraît indispensable, c'est le respect à l'égard du lecteur. Un travail doit dire et montrer comment il est fait. C'est à cette condition qu'il peut non

seulement ne pas être trompeur, mais être positivement utile. Tout livre dessine autour de lui un champ de travail virtuel et il est jusqu'à un certain point responsable de ce qu'il rend possible ou impossible. Un livre – je parle bien sûr de ces ouvrages de savoir – qui brouille ses manières de faire n'est pas quelque chose de très bien. Je rêve de livres qui seraient assez clairs sur leur propre manière de faire pour que d'autres puissent s'en servir librement, mais sans chercher non plus à brouiller les sources. La liberté d'usage et la transparence technique sont liées.

(*Dits et écrits* 414)

Quelques années auparavant, à Raymond Bellour, il souligne le fait que, dans sa méthode d'analyse des documents, tout devient source historique : « Il ne doit pas y avoir de choix privilégié. Il faut pouvoir tout lire, connaître toutes les institutions et toutes les pratiques. [...] Ce qui fait qu'on traitera dans la même foulée *Don Quichotte*, Descartes et un décret sur la création des maisons d'internement par Pomponne de Bellièvre » (Bellour 17).

En regard de cette manière de présenter ses propres recherches, la possibilité est aujourd'hui offerte de reprendre cette question des pratiques réelles de Foucault à partir des nouvelles méthodologies de recherche et d'enquête qui se sont développées autour des archives du travail intellectuel et qui ont été insufflées par Bruno Latour, Roger Chartier (cf Kraus) et, plus récemment, Christian Jacob et sa notion de pratique savante qui lui permet de regrouper sous cette notion opératoire l'ensemble des opérations manuelles, discursives et intellectuelles mobilisées dans la production ou la réception d'un savoir.

Dans son célèbre article « Les Vues de l'esprit, une introduction à l'anthropologie des sciences et des techniques », et à propos des laboratoires scientifiques Bruno Latour donne en détail le programme de cette anthropologie du savoir dans laquelle s'inscrit cette recherche :

Pourquoi les inscriptions de toutes sortes sont-elles aussi importantes pour les chercheurs, les ingénieurs, les architectes, tous ceux qui pensent avec leurs yeux et leurs mains ? Parce qu'elles offrent un avantage unique lors des discussions : « Vous doutez de ce que je dis ? Vous allez voir, je vais vous montrer ! » et sans remuer de plus de quelques centimètres, l'orateur déploie devant les yeux de ses critiques autant de figures, diagrammes, planches, silhouettes qu'il en faudra pour convaincre. Aussi médiates que soient ces inscriptions, aussi lointaines que soient les choses dont on parle, des chemins à double voie s'établissent. L'objecteur se trouve dominé par le nombre de choses dont parle l'orateur, toutes présentes dans la salle. Il peut douter de chacune d'elles, mais toutes ensemble, elles composent une formidable preuve. Nous sommes tellement

habitué à recourir à ces alliés ; que nous avons oublié ce que c'est que penser sans index, sans bibliographies, sans dictionnaires, sans fiches bristol, sans physiographes, sans cartes, sans diagrammes [...]. (87)

Plus récemment, l'historien Christian Jacob a voulu, en insistant sur la matérialité du travail intellectuel, redonner une nouvelle direction à ce même programme :

[S]i un anthropologue pénètre dans le bureau d'un universitaire contemporain ou dans la salle d'une soutenance de thèse, ou encore dans une bibliothèque ou une salle de colloque, sans préjuger du contenu et du sens des activités qui y prennent place, il sera amené à observer des gestes, des déplacements, des postures, une répartition spatiale des acteurs, un mobilier, des ordinateurs, des étagères avec des livres dont la répartition est signifiante, etc. S'il se penche sur le bureau d'un chercheur, il remarquera des pages imprimées, des fiches, un livre ouvert, différents crayons et stylos, une agrafeuse, un écran et un clavier d'ordinateur, des post-it, [...] des croquis, des paragraphes de texte, une carte, un tableau, des statistiques. [...] Notre hypothèse de travail est que ces différentes échelles d'observation, de l'environnement architectural à l'écran d'ordinateur, nous apprennent quelque chose de signifiant sur la nature du travail savant, au sens de séquences de gestes et d'opérations où le mental est relié au matériel, par le biais d'instruments, d'objets et d'inscriptions.

(Müller 122)

Il y a tout intérêt à ouvrir un terrain d'enquête sur les traces de Michel Foucault en bibliothèque qui informerait, de manière inédite, sur les procédures de lecture qui étaient les siennes. Autrement dit, analyser ces gestes minimes, répétés quotidiennement, qui font le métier d'intellectuel dans la deuxième moitié du XXe siècle. Grâce à Daniel Defert, l'on sait que la chaîne d'écriture de Foucault était composée d'au moins trois moments : « d'abord la version de ce qu'il ne fallait pas dire, qu'on pensait un peu spontanément. Ensuite, il y avait la reprise de tout ça à partir d'un travail de recherches, ce qui demandait facilement trois ans. Une fois ce travail de recherches fait, il y avait une réécriture » (Bellon 4). Dans cette enquête sur les « inscriptions » multiples de son activité de philosophe, le cas des *Mots et les choses* qu'il publia dans la collection de la « Bibliothèque des sciences humaines » chez Gallimard en 1966 est exemplaire de sa pratique.

En effet, le dossier préparatoire des *Mots et des Choses* que Daniel Defert¹ a mis temporairement à notre disposition, révèle que Michel Foucault rédigea en amont plusieurs centaines de fiches de prises de notes. Ces fiches témoignent à la fois chez

¹ Qu'il nous soit permis de remercier ici Daniel Defert de nous avoir offert la possibilité de consulter ces documents.

lui, d'une manière de cerner un domaine (comme la grammaire générale, l'histoire naturelle, l'économie...), de baliser un champ de recherche et d'organiser l'ordre du travail. L'existence de ces fiches n'est pas sans rapport non plus avec la méthode « archéologique » que Foucault met alors en place et qui doit lui permettre d'identifier les concepts par les connexions qui régissent leur emploi. Cette analyse, ajoute-t-il, a pour but de « reconstituer le système général de pensée dont le réseau, en sa positivité, rend possible un jeu d'opinions simultanées et apparemment contradictoires » (*Les Mots et les choses* 89–90).

Bien que notre approche se rapproche de l'anthropologie des pratiques culturelles et savantes et que ce travail d'investigation ait pour matériau principal un objet qui l'exclut à priori – si on s'en tient à des définitions strictes – d'une approche codicologique (codex, parchemins vs papier du XX^e s.) ou paléographique (époque), nous avons voulu conserver, au vu de la nature du dossier, certaines techniques d'analyse matérielle propre à l'analyse des manuscrits modernes (Bustarret 47–60) mais surtout des techniques de l'archivistique pour reconstituer certaines habitudes d'écriture et de travail du philosophe ; comme, par exemple, sa manière d'assembler ses matériaux en liasses ou, au contraire, de les laisser sous forme de feuilles volantes.

Nous avons entrepris l'analyse matérielle de cet ensemble de documents, de ses découpages (dossiers, sous-dossiers, listes bibliographiques, fiches,...), ainsi qu'une description systématique de chacun des feuillets (rature, soulignement, organisation de la fiche) en procédant à un inventaire XML, qui, une fois complet, renseignera sur les usages d'une méthode de travail explicite. Par ailleurs, ce travail d'inventaire nous a obligé à mieux renseigner les circonstances de la transmission de ces documents et à analyser s'ils avaient fait, ou non, l'objet d'interventions ultérieures.

Au delà de l'approche disciplinaire, l'objet de cet article sera de partager avec les spécialistes de la codicologie et de la paléographie, des outils et méthodes du numérique qui nous ont permis de :

- numériser l'ensemble des documents manuscrits qui composent ce dossier,
- créer un catalogue informatisé de notices descriptives de ces fiches
- et, enfin, de concevoir une véritable bibliothèque virtuelle qui réunira à la fois ce catalogue numérique et l'ensemble des fiches numérisées.

Du protocole de numérisation au Langage de Modélisation Unifiée (Unified Markup Language ou UML) pour modéliser le site web final, en passant par la Description Archivistique Encodée en XML (Encoded Archival Description ou EAD)² des fiches manuscrites assistée par le logiciel de gestion d'archives, The Archivists' Toolkit (AT), ces outils et méthodes méritent d'être décrits, car c'est grâce à eux que ce corpus qui

² Pour plus de détails on pourra consulter le dictionnaire de balises EAD. C'est également le format qu'a choisi avant nous l'Institut Jean-Toussaint Desanti pour décrire les manuscrits du philosophe Jean Toussaint Desanti (L'archive numérique).

donne à voir la « bibliothèque » choisie par le philosophe pour la rédaction des *Mots et les choses* est aujourd'hui intelligible et surtout utilisable pour la recherche.

1. Les manuscrits du travail philosophique : contexte et réception de l'archive

1.1. Contexte

Dans une récente interview, Daniel Defert, compagnon du philosophe, relate la manière dont Foucault détruisait les manuscrits de ses ouvrages au fur et à mesure qu'ils étaient publiés :

J'ai cédé à la BNF le seul manuscrit qui restait, de *l'Histoire de la sexualité*. Le livre est paru quand Foucault était déjà à l'hôpital, et il n'a donc pas pu détruire le manuscrit. Sans quoi, il l'aurait détruit comme tous les autres : je l'ai vu jeter ses manuscrits, et je ne lui ai jamais dit qu'il devrait les conserver : Foucault aurait rigolé, et les aurait détruits encore plus vite.

(Bellon 4)

Parmi les documents qui échappèrent à cette destruction systématique et qui constituaient en somme une archive vivante de ses recherches, les notes venant du travail de Foucault en bibliothèque permettent non seulement de préciser l'importance de la lecture dans le processus de compréhension et de re-création d'un appareil notionnel ainsi que dans le processus d'écriture du philosophe, mais aussi certaines méthodes de travail rigoureuses. Daniel Defert d'ajouter :

Foucault avait en effet une manière de prendre des notes très spéciale : une citation par page, qu'il rangeait thématiquement. Or, corréler ces références au livre qu'on publie est pratiquement impossible. Les éditeurs [...] sont souvent obligés de retourner en bibliothèque, à la recherche des références de Foucault [...]. L'essentiel des archives que Foucault a laissé, ce sont ces citations, peut-être dix mille feuilles au format A4. On n'a pas forcément idée de cet énorme travail d'archive de Foucault, mais il lisait très vite, énormément : toute sa vie durant, Foucault a passé six heures par jour en bibliothèque [...].

(Bellon 5)

On sait par les différents biographes de Foucault, Didier Eribon et David Macey, que celui-ci était un des grands habitués de la rue Richelieu, mais la Bibliothèque nationale de France ne fut pas l'unique ressource du philosophe. À partir des années 1980, il se rendit également à la Bibliothèque dominicaine du Saulchoir, rue de la Glacière dans

le treizième arrondissement de Paris. Foucault a fréquenté aussi d'autres bibliothèques comme la *Carolina Rediviva* d'Uppsala ou celle de Varsovie. Dans cette liste, il en est une peut-être plus décisive, celle de sa formation, à l'Ecole normale supérieure de la rue d'Ulm où, dès les années 1900, les élèves et professeurs pouvaient naviguer à leur guise, passant d'un ouvrage d'histoire des sciences à un volume de poésie. On ignorait également que Foucault avait porté de l'intérêt à certains ensembles de manuscrits et d'archives conservés dans d'autres bibliothèques parisiennes. L'analyse de l'un des dossiers préparatoires à l'*Histoire de la folie*, relatif à ses recherches en bibliothèque, permet d'apporter un éclairage inédit sur les pratiques informatives de Foucault, sur ses différents « savoirs-faire » ou « techniques intellectuelles » qui lui sont propres au cours de la rédaction de sa thèse. On trouve en effet dans ce dossier³ la trace de ses consultations dans différentes bibliothèques parisiennes dont la Bibliothèque historique de la ville de Paris, la Bibliothèque nationale de France (collections Clairambault et Joly de Fleury) mais aussi la Bibliothèque du duc de la Vallière, ou encore celle de Sainte Geneviève.

1.2. Réception des pièces manuscrites

Pour *Les Mots et les choses*, c'est l'ensemble du dossier préparatoire qui a été mis à disposition de notre équipe sous forme de 1712 images.tiff (soit 856 fiches numériser recto/verso), en résolution 300 DPI, taille moyenne 2550px L. * 3600px H. Ces fiches sont réparties dans cinq dossiers titrés qui, pour une large mesure, reprennent les grandes thématiques de l'ouvrage : « Analyse des richesses » (176 fiches) ; « Grammaire » (230 fiches) ; « Homme » (18 fiches) ; « Langage » (151 fiches) et « Histoire naturelle » (281 fiches). Ces fiches sont de trois types : fiches de prises de notes qui ont une disposition identique (voir Partie 2), fiches bibliographiques où le philosophe barre les titres des ouvrages vus, et fiches thématiques, plus rares, qui lui permettent de réunir sous un mot commun diverses lectures.

L'équipe de recherche en charge du dossier a entrepris, en collaboration étroite avec l'Ecole normale supérieure de Lyon pour la partie technique,⁴ de constituer un catalogue systématique, critique et philologique⁵ de cette archive pour en conserver et en préserver

³ Ce dossier, encore en court de traitement, fait parti des notes préparatoires de Foucault. Il est relatif aux travaux bibliographique du philosophe dans plusieurs bibliothèques parisiennes, dont la Bibliothèque Nationale de France, la bibliothèque de l'Arsenal, mais aussi la bibliothèque Sainte-Geneviève et la bibliothèque historique de la ville de Paris. Un autre dossier préparatoire de l'*Histoire de la folie* est présenté sur le portail Foucault de l'IMEC.

⁴ Le projet de *La Bibliothèque Foucauldienne* a été rendu possible grâce au financement de l'Agence Nationale pour la Recherche. Il comprend les équipes suivantes : l'équipe Anthropologie de l'écriture (IIAC – EHESS – Paris), resp. Philippe Artières et l'équipe Philosophie de la politique (UMR 5206 – ENS de Lyon), resp. Jean-Claude Zancarani.

⁵ Il ne s'agit pas de chercher l'origine de chacun des travaux du philosophe, ni même de travailler sur ses brouillons, comme pourraient le faire les généticiens de la littérature mais nous intéresser à l'ensemble des

l'existence mais surtout pour étudier l'appareil notionnel et les méthodes de travail de Michel Foucault sur les sources utilisées.

Tout d'abord, grâce à la description des fiches et à leur comparaison avec les textes sources lus par Michel Foucault, il s'est agi de comprendre et d'analyser le travail existant en amont du processus d'écriture (travail bibliographique, lectures, modes de consultation des sources).

Ensuite, grâce à une comparaison à grande échelle entre ces fiches et le texte produit par Michel Foucault, il a été question de revenir sur les modes d'extractions de Foucault (comment le philosophe passe du « nomadisme » bibliographique à un travail méticuleux de commentaire de texte), sur les pratiques de restitutions des sources (quelle pratique de la bibliographie ? quelle pratique de la citation ?), et sur la construction de l'appareil notionnel et la création de concept.

Ces fiches n'ont pas fait l'objet d'une transcription systématique ni d'une réédition au format texte mais d'un inventaire détaillé, augmenté d'un index, de listes bibliographiques, de descriptions, d'explications comparatives entre les textes lus et les fiches, et entre les fiches et les textes produits. Nous allons maintenant décrire les outils et méthodes qui nous ont aidés à réaliser ce travail et à le partager.

2. Une archive numérique : outils et méthodes

Notre projet tend à un double objectif : conserver une copie numérique et intellectuelle de l'archive et créer un instrument de recherche, capable d'interroger l'ensemble de l'archive privée. Les papiers de travail de Michel Foucault ont donc fait l'objet d'une description systématique, plutôt que d'une transcription.

2.1. Pourquoi avoir décrit aussi finement l'archive ?

Le dossier physique *Les Mots et les choses* n'étant pas encore confié pour conservation à l'Institut Mémoires de l'Édition Contemporaine (IMEC), il pourrait être détruit à tout moment. Nous nous devons d'en garder une description archivistique poussée ; en nous appuyant sur sa reproduction numérique. Or l'archivistique, a, depuis les années 1960, bénéficié, au même titre que la bibliothéconomie et la philologie, des évolutions du numérique. En effet, la création des modèles structurés d'échange et de traitement bibliographique, les formats Machine Readable Cataloging (MARC) en 1965, Universal Machine Readable Cataloging (UNIMARC ; Willer) en 1977, le standard Text Encoding Initiative (TEI ; Burnard) en 1987 qui, lui, propose des guidelines de méthode d'encodage pour la mise en ligne de texte, et enfin le format EAD en 1993, inspiré de la TEI et des

traces qui témoignent de ses pratiques de lecture, d'enquêter sur ses prises de note en bibliothèque, sur ses manières d'archiver, mais aussi sur la fonction citationnelle, les notes de bas de page (Grafton), l'usage de la bibliographie, et l'ensemble du paratexte (Genette) – sans négliger l'examen de sa propre bibliothèque.

formats MARC, sont autant d'évolutions qui permettent aujourd'hui de répondre aux besoins de notre projet.

2.2. Encodage : EAD vs TEI msDesc

Arrêtons-nous ici sur les raisons qui nous ont fait choisir le standard adapté à l'archivistique, EAD, plutôt que le standard classique de la philologie numérique⁶, le standard TEI⁷.

En effet, le projet Européen mené par Peter Robinson entre 1999 et 2001 *Manuscript Access through Standards for Electronic Records* (MASTER) offre lui aussi un modèle TEI de description de sources primaires manuscrites très intéressant : *TEI Manuscript Description* ou *TEI msDesc*. Pour choisir entre TEI msDesc et EAD, nous avons pris en compte trois critères.

Un encodage pour tous les niveaux de l'archive

Tout d'abord le modèle msDesc, s'il prévoit un niveau de description très fin d'une source primaire, ne rend pas bien compte de l'ensemble dans lequel la source existe. En effet, le niveau de description le plus haut prévu par ce modèle est l'élément <msDesc/>, ce qui correspondrait à notre niveau le plus bas, <c level="item"> en EAD. Ce modèle prévoit néanmoins la possibilité de décrire une collection de pièces manuscrites, en intégrant le niveau <msDesc> dans l'élément bibliographique <listBibl> : « However, in cases where the document being encoded is essentially a collection of manuscript descriptions, the msDesc element may be used in the same way as the bibliographic elements (bibl, biblFull, and biblStruct) making up the TEI element class model.biblLike. These typically appear within the listBibl element » (10.2).

Avec ce modèle, la description de l'archive serait placée dans l'en-tête du fichier, avec une transcription ou des images dans le corps du fichier XML :

```
<TEI.2>
  <teiHeader>
    <fileDesc>
      <sourceDesc>
        <msDesc/>
      </sourceDesc>
```

⁶ En nous appuyant sur les définitions de la philologie numérique de Franck Neveu (226–27) et François Rastier (73), nous entendons ici, par philologie numérique, l'étude informatisée des textes manuscrits qui nous permettront de découvrir les processus et procédés d'écriture du philosophe (ici, ceux de Michel Foucault) : des lectures à l'appropriation de champs disciplinaires et notionnels ; de l'appropriation à la rédaction, comment les manuscrits laissés par le philosophe nous permettent de comprendre comment il écrivait. Cette étude est dite informatisée à deux titres : premièrement les manuscrits et leur description sont annotés/étiquetés pour permettre une automatisation des recherches. D'autre part, l'application finale qui permettra la visualisation des résultats devra être modélisée afin d'être développée et/ou optimisée.

⁷ Initiative à l'origine de la création de guides de méthodes d'encodage de textes pour les sciences humaines et sociales.


```

    </fileDesc>
  </teiHeader>
  <text>
    <pb facs="image1" />
  </text>
</TEI.2>

```

Or, si nous considérons notre archive, c'est bien l'ensemble des pièces qui doit être décrit et interrogeable dans son organisation originale. On doit pouvoir interroger n'importe quel niveau de l'ensemble, soit au moins six niveaux : le fonds, les groupes de documents, les séries, les dossiers, les sous-dossiers, et les pièces. En adaptant les éléments <listBibl/>, corpus et <msDesc/>, nous y serions parvenu, mais nous avons, pour des raisons de temps impartis très courts, choisi d'utiliser le format EAD qui offrait un jeu d'éléments et d'attributs adaptés à l'organisation et la nature sérielle de notre archive. Nous reviendrons sur le détail des balises EAD choisies pour notre description, dans la partie 2.3.

Des outils plus pratiques

Voyons maintenant notre deuxième critère : EAD nous offre une palette d'outils facilement utilisables et paramétrables que n'offre pas encore la TEI. Les logiciels de gestion d'archives permettent de renseigner et décrire l'archive dans des interfaces plus accessibles aux chercheurs en sciences humaines et sociales que des éditeurs de code : Digitool, ICA-Atom, The Archivists' Toolkit (AT), Archon (Spiro) etc. Certains de ces logiciels offrent également la possibilité de publier sur le web l'archive interrogeable (index de personnes, index thématique, recherche plein texte). Suite à des tests d'exports XML très propres et satisfaisants, nous avons donc finalement opté pour l'Archivists' Toolkit. C'est une application JAVA Client/Serveur élaborée par l'*University of California San Diego Libraries*, la *New York University Libraries* et la *Five Colleges, Inc. Libraries*. Ce projet a été financé par l'*Andrew W. Mellon Foundation*. Régulièrement, l'archive est exportée au format EAD XML pour être corrigée ou adaptée par les ingénieurs qui utilisent l'éditeur de code oXygen.

Correspondance avec les formats bibliographiques et les normes archivistiques

Les logiciels de gestion d'archives assurent la correspondance entre EAD et le format MARC ; et entre EAD et la Norme Générale Internationale de description archivistique ISAD(G). En effet les correspondances entre ces différents formats font partie du processus de création et de mise à jour de la DTD 2002 de l'EAD.

Pour ces 3 raisons notre choix s'est donc porté sur EAD. La TEI sera utilisée, au cas par cas, quand un membre de l'équipe décidera de transcrire telle ou telle pièce : les transcriptions seront enregistrées et stockées comme des objets numériques (au même titre que des images) : une simple déclaration et pointeur dans l'élément <dao/> du fichier TEI suffiront, comme nous le verrons plus loin.

Ainsi, le format et les outils choisis nous ont donné la possibilité de décrire l'archive pièce par pièce, de rendre compte des regroupements (thématiques, typologiques, etc.) effectués par Foucault et de les déplier, de rendre compte du découpage et de l'organisation interne d'une fiche, de créer des index et de repérer dans un document des termes ou notions clefs, c'est à dire de donner à voir le contenu notionnel des sources d'après la lecture qu'en a fait Foucault.

2.3. EAD et le choix des niveaux de description

En respectant le dictionnaire de balises EAD, la DTD 2002 et le manuel d'encodage publiés respectivement en 2002 et 2005 par le groupe CG46/CN357/GE3⁸ de l'Agence Française de Normalisation (AFNOR), nous avons choisi les niveaux de description qui pouvaient le mieux représenter l'archive existante, tout en laissant ouverte la possibilité de tout type d'ajouts ultérieurs et d'ajouter des niveaux de description ultérieurs.

En voici une description accompagnées d'illustration. Nous retenons donc pour l'instant les niveaux suivants :

- Fonds : Les papiers de travail de Michel Foucault
- Recordgp ou Sous-fonds : *Les mots et les choses*
- Series : Les fiches de lecture
- File : Analyse des richesses, etc.

Le fonds <archdesc level="fonds">

Les papiers de travail de Foucault (séparés entre plusieurs lieux dont la BNF, l'appartement de Daniel Defert, celui d'Arlette Farge, l'IMEC, etc.) pourraient faire considérer l'ensemble comme une archive artificielle et *non-organique*. Or, c'est le processus de production des pièces qu'il faut considérer, et non leur localisation, pour juger si l'ensemble est *organique* ou *non-organique*.

En EAD, on considère comme *organique* un ensemble de pièces homogènes au moment de la production, par exemple un ensemble de pièces produites par un même auteur, ou une même institution. Les pièces à décrire ici, « *les papiers de travail de Michel Foucault* », forment un ensemble organique au moment de leur production, puisqu'ils proviennent du même auteur : nous avons donc choisi comme niveau de description le plus haut : le *fonds*. Une *collection* en revanche, serait, par exemple, un ensemble de pièces disparates (différents types de pièces ou des pièces produites par différents producteurs, etc.) réunies artificiellement autour d'un sujet.

⁸ Parmi ce groupe de travail AFNOR, Denise Ogilvie (Archives Nationales de France) nous a accordé un entretien très utile à ce sujet.

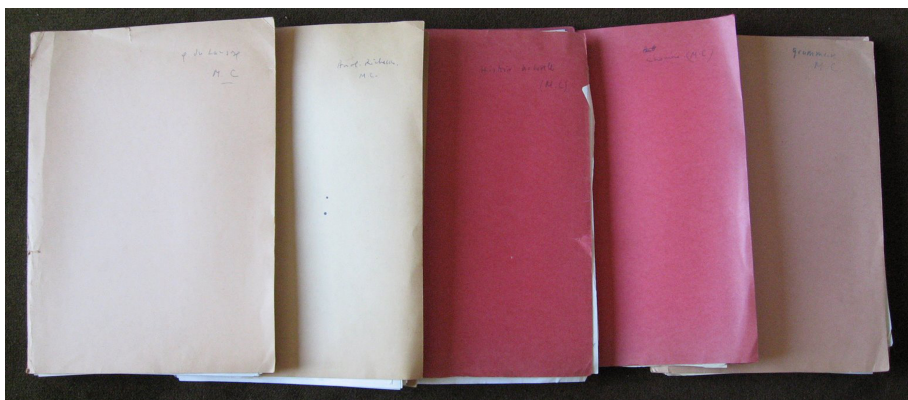


FIGURE 1. Ces 5 dossiers correspondent aux grands chapitres de *Les Mots et les Choses* : « Philosophie du langage », « Analyse des Richesses », « Histoire Naturelle », « Homme », « Grammaire ».

Le groupe de documents <c level="recordgrp">

Les Mots et les choses représente un groupe de documents appartenant au fonds Papiers de Michel Foucault. Il contient tous types de papiers de travail ayant été utilisés par Michel Foucault pour rédiger *Les Mots et les choses*. La raison pour laquelle le niveau de description *Les mots et les choses* a été choisi comme englobant les fiches de lecture sera détaillée plus bas, avec la description des « dossiers ».

Les séries <c level="series">

En EAD, le terme *series* s'utilise pour des pièces qui sont produites à la chaîne (par exemple un formulaire type produit en série par une institution). Il apparaît donc logique d'utiliser *series* pour désigner l'ensemble des fiches de lecture et non plus pour désigner l'ensemble *Les mots et les choses*.

Les dossiers <c level="file">

Nous avons choisi *Les mots et les choses* comme un groupe de documents englobant les fiches de lecture contenues dans les 5 dossiers (« Philosophie du langage », « Analyse des Richesses », « Histoire Naturelle », « Homme », « Grammaire ») (cf. fig. 1) qui correspondent aux thématiques annoncées dans la préface de l'ouvrage : « pour former le socle positif des connaissances telles qu'elles se déploient dans la *grammaire* et dans la *philologie*, dans l'*histoire naturelle* et dans la biologie, dans l'*étude des richesses* et dans l'économie politique. [...] Mais à mesure que les choses s'enroulent sur elles-mêmes, ne demandant qu'à leur devenir le principe de leur intelligibilité et abandonnant l'espace

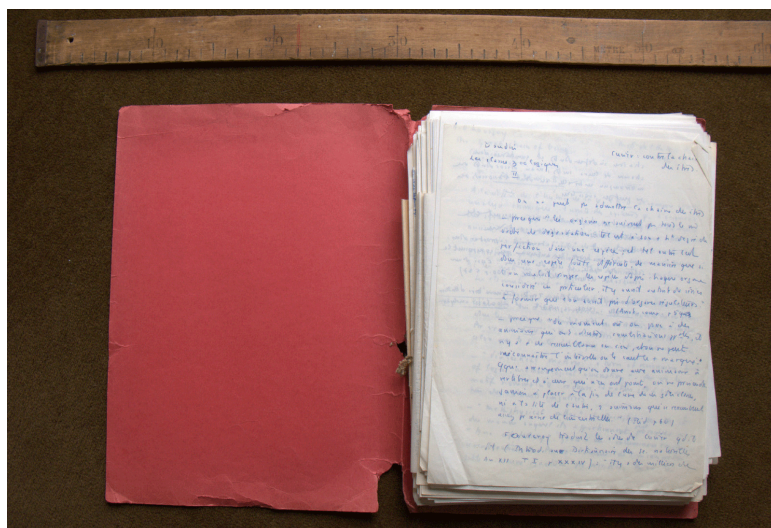


FIGURE 2. Le dossier « Histoire Naturelle ».

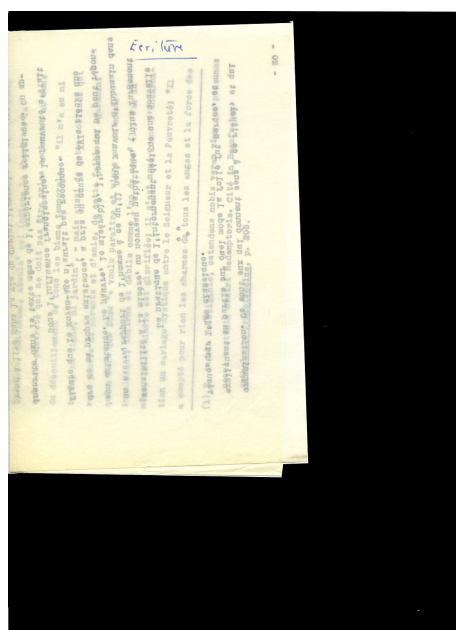


FIGURE 3. Page A4 pliée en deux : un sous-dossier présent dans le dossier « Grammaire » : « Écriture ».

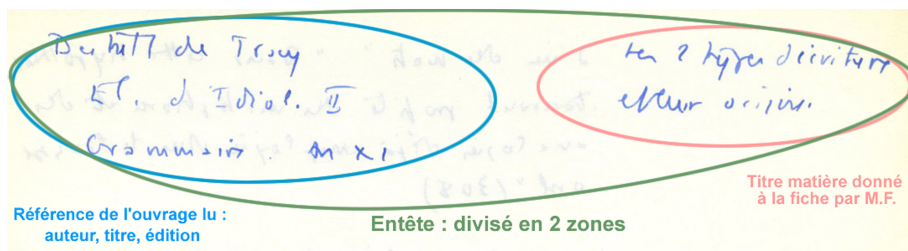


FIGURE 6. En-tête d'une fiche de lecture (1).

de la représentation, l'*homme* à son tour entre, et pour la première fois, dans le champ du savoir occidental » (13–15).

En consultant le sommaire, le rapprochement est encore plus net et permet de déduire les correspondances suivantes :

Le chapitre un, *La prose du monde* correspond au dossier « Grammaire » ; le chapitre deux, *Parler* correspond au dossier « Philosophie du langage » ; le chapitre trois, *Classer* correspond au dossier « Histoire naturelle » ; le chapitre quatre *Échanger* correspond au dossier « Étude des richesses /économie » ; le chapitre cinq, *L'homme et ses doubles* correspond au dossier « Homme ».

Ces sous-ensembles sont considérés comme des *dossiers* (<c level="file">) qui sont englobés dans une *série* (<c level="series">) de *fiches de lectures*, qui sont elles-mêmes englobées dans un *ensemble de documents* (<c level="recordgrp">) ayant servi à la rédaction de l'ouvrage *Les Mots et les choses*.

Les sous-dossiers <c level="file">

Les dossiers sont divisés en plusieurs sous-dossiers, que nous considérerons comme des *dossiers* (<c level="file">).

La pièce = une fiche de lecture <c level="item">

Le niveau de description le plus bas (celui qui va demander le plus de minutie et d'attention), est celui de la pièce : la fiche de lecture.

Foucault consacre une page recto-verso pour citer un passage d'un ouvrage qu'il cite sous un titre matière. Cette pratique permet à Michel Foucault de découper l'ouvrage à partir de ses propres problématiques. Pour chaque fiche, Foucault nous donne au moins :

- le nom de l'auteur ;
- le titre de l'ouvrage et l'année de publication (souvent la première édition) ;
- un titre matière qui lui permet de regrouper les passages cités.

En voici un premier exemple sur la figure 6. En voici un deuxième, grâce à la figure 7.

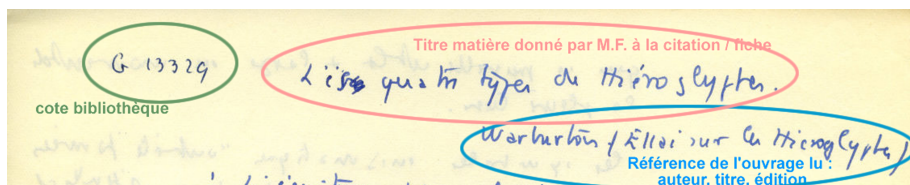


FIGURE 7. En-tête d'une fiche de lecture (2).

La version 2002 de la DTD EAD nous offre en effet la possibilité de renseigner trois niveaux de description des fiches : un niveau général regroupant le titre de la fiche, la langue, le numéro d'inventaire des documents quand celui-ci est renseigné par Foucault ; un niveau biographique et thématique qui permet de renseigner le nom de l'auteur et les différents thèmes abordés par Foucault dans chacune des fiches — il a fallu pour cela élaborer un vocabulaire thématique à partir des fiches ; enfin un niveau bibliographique et analytique qui permet de renseigner la référence de l'ouvrage cité, et d'analyser le document à partir de plusieurs points d'entrée (mots soulignés, expression traduite par Foucault, pagination des citations). Chaque fiche a aussi fait l'objet d'une description matérielle avancée dans laquelle sont renseignées la nature du support et la graphie. Voici comment nous récupérerons ces diverses informations :

par la balise de titre : <unittitle></unittitle> peut elle-même contenir plusieurs balises de titre. On peut donc utiliser pour décrire un des deux cas de figure ci-dessus : <unittitle>Destutt de Tracy. Element d'Idéologie II. Grammaire - XI. <title type="TM_MF">Les deux types d'écriture et leur origine</title></unittitle>.

par l'indexation : pour créer les index de « mots sujets », ou « titres matières », plutôt qu'utiliser le langage d'indexation Rameau, nous avons trouvé plus intéressant d'utiliser le vocabulaire de Michel Foucault pour donner à voir la construction d'un appareil notionnel : cela nous permettra d'indexer le contenu propre au corpus. A cet effet, on indexera, dans les fiches, les titres matières (à gauche souvent, dans la zone de titre), et les mots mis en évidence par Michel Foucault.

2.4. Fichier EAD et objets numériques

Il est intéressant d'observer dans l'encodage EAD XML (fig. 8) d'une pièce comment les objets (livres de référence lus par Michel Foucault, fiches de lecture, ouvrages produits par Michel Foucault, images obtenues par la numérisation) sont liés au fichier de

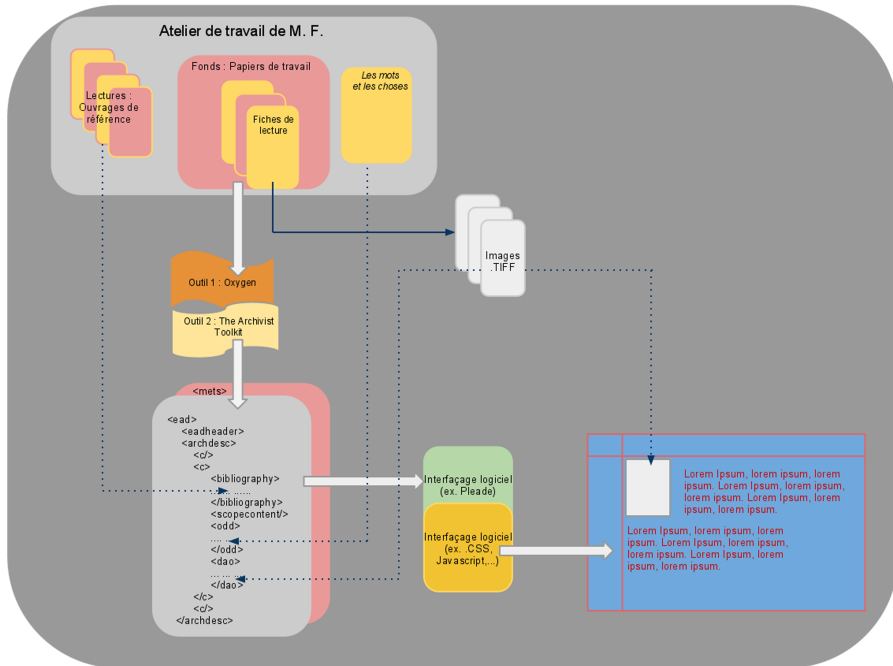


FIGURE 8. Un exemple simplifié de la description d'une pièce (<c level="item">).

description EAD XML.

Dans la balise <bibliography/> ce sont toutes les références nécessaires pour retrouver la littérature à l'origine des fiches que l'encodeur spécifiera.

Dans la balise <odd/> ce sont les ouvrages et les passages produits par Michel Foucault grâce à la fiche de lecture qui seront spécifiés. Enfin, les liens vers les différentes versions de la fiche, c'est-à-dire ici vers les images numérisées correspondantes (mais il pourrait tout aussi bien s'agir d'une transcription TEI), seront faits dans le balise <dao/> : l'objet archivistique numérique. Nous envisageons d'utiliser cette balise pour les fiches que les chercheurs décideront de transcrire en TEI.

2.5. L'application finale pour présenter les résultats : modélisation UML

Les outils utilisés jusqu'ici ne suffisent pas à rendre ergonomique, interactive et hautement intuitive l'application finale, ni à définir pour cette même application, les


```

<c id="ref43" level="item">.
  <did>.
    <umittitle>Destutt de Tracy. Element d'Idéologie II. Grammaire -- XI. .
      <title type="TM_MF">Les deux types d'écriture et leur origine</title></umittitle>.
      <unitid>0007</unitid>. TM_MF = Titre matière donné par Michel Foucault
      <langmaterial>.
      <language langcode="fre"/>.
      </langmaterial>.
      <physdesc>.
      <extent>2.0 digitized images</extent>.
      </physdesc>.
    </did>.
    <bibliography id="ref45">.
      <head>Référence de l'ouvrage</head>.
      <bibref>Destutt de Tracy, Eléments d'idéologie. Seconde partie. Grammaire, Paris : Courcier, an XI-1803.
    </bibliography>.
    <scopecontent id="ref46">.
      <head>Analyse du document</head>.
      <p>Mot(s) souligné(s) :</p>.
      <p>Michel Foucault nous donne ici la cote de l'ouvrage emprunté pour la réalisation de cette fiche :</p>.
      <p>Les passages cités par Foucault dans cette fiche sont issues des pages : p. 307, p. 308, p.309 et p. 3
    </scopecontent>.
    <odd id="ref47">.
      <head>Commentaires</head>.
      <p>Traduit par Foucault de la langue originale :</p>.
      <p>Michel Foucault cite, dans <i>Les mots et les choses</i>, l'auteur faisant l'objet de cette fiche : la
      <p>Cet auteur est cité dans cet autre ouvrage de Michel Foucault :</p>.
    </odd>.
    <dao ns2:actuate="onRequest" ns2:show="embed" ns2:title="page 01 - recto" ns2:href="PMF_MC_FL_Gr_Ec_HR_003" .
      <daodesc>.
      <p>page 01 - recto (PMF_MC_FL_Gr_Ec_HR_003)</p>.
      </daodesc>.
    </dao>.
    <controlaccess>.
      <persname source="ingest">Destutt de Tracy, Antoine-Louis-Claude (1754-1836)</persname>
      <subject source="VNF">hiéroglyphes</subject>.
      <subject source="VMF">origine de l'écriture</subject>.
      <subject source="VMF">types d'écriture</subject>.
      <subject source="VMF">égyptien ancien (langue) -- écriture hiéroglyphique</subject>.
    </controlaccess>.
  </c>.

```

FIGURE 9. Diagramme de navigation pour l'internaute invité.

différents niveaux d'accès aux reproductions iconographiques, dans le respect de la volonté des légataires. Nous venons donc d'entrer dans une nouvelle phase du projet, celle de la modélisation UML de l'application web finale. Cette modélisation s'appuie sur la démarche définie par Pascal Roques dans le premier chapitre de son ouvrage (2-21). Elle nous permettra de définir les acteurs, les cas d'utilisation, la maquette, ainsi que tous les diagrammes (les diagrammes de séquences système, de classes participantes, des interactions, de navigation, et de classes de conception) nécessaires au choix d'un logiciel (Pleade)⁹, ou au développement d'une application.

Ces diagrammes nous permettront d'envisager l'application finale, jusqu'au détail le plus fin. On verra à titre d'exemple le diagramme de navigation d'un internaute

⁹ Exemples de logiciels envisagés pour la mise en ligne : Pleade (AJLSM 2009).

« invité », c'est à dire de l'internaute bénéficiant d'identifiants de connexion pour une navigation simple avec un accès aux textes et aux images (cf. fig. 9). Il nous faudra définir ainsi les diagrammes nécessaires pour chaque type d'utilisateur, et les synthétiser jusqu'à obtenir un modèle de conception détaillé.

Conclusion

Les apports du numérique ont été essentiels à ce travail de recherche sur les papiers de travail de Michel Foucault. En effet, ni l'annotation assistée de ses fiches de lecture manuscrites, ni l'automatisation des recherches dans l'archive et les objets qui lui sont liés (références d'ouvrages, ouvrages produits, images numérisées), ni la modélisation de l'application finale qui offrira un pan de la bibliothèque foucaldienne aux yeux des internautes, n'auraient été possibles sans les apports méthodiques et logiciels de l'informatique à la philologie du XXI^{ème} siècle. Pourtant, ces « apports » ont également été sources de questionnements difficiles, comme le choix du format EAD plutôt que le standard TEI. Pour cette question précise, c'est, comme nous l'avons vu plus haut, EAD, une norme compatible aux formats bibliographiques, aux normes archivistiques et permettant une exploitation logicielle plus adaptée et simplifiée qui l'a emporté.

Dans l'état actuel d'avancement du travail de description, le traitement archivistique de ces fiches nous a permis d'envisager de nouvelles hypothèses de travail autour du travail foucaldien et tout d'abord de reprendre la question du renversement des hiérarchies et le choix délibéré de Foucault de choisir certains auteurs dits « secondaires » pour appréhender des champs de savoir parfois nouveaux pour lui (exemple : économie).

Nous avons également pu mieux comprendre le degré des citations. En effet, Foucault cite peu par rapport à ce qu'il a lu et surtout noté. Les exemples sont nombreux d'auteurs cités dans plusieurs fiches mais qui n'apparaissent pas dans l'ouvrage publié (c'est le cas par exemple du révérend père Coeurdoux connu pour son analyse de la ressemblance entre le sanscrit et le latin. Si son nom apparaît dans 4 fiches, Foucault ne le cite qu'en page 305, dans le corps du texte sans en dire plus).

Au-delà de ces premières hypothèses concernant le travail de Foucault, d'autres hypothèses sont envisageables concernant cette fois-ci le travail intellectuel en général. Foucault est pour cela un « laboratoire » de première importance. Ce type de travail sur les archives de la recherche permettrait par exemple de proposer une approche inédite du travail philosophique au XX^e siècle qui rendrait compte des principes de transformations de l'attitude rhétorique des intellectuels dans leur manière de présenter leurs idées, de les mettre en œuvre et de les diffuser : à l'université, devant un public cultivé, ou à la radio, à la télévision. Ce travail sur les archives permet aussi de mieux comprendre la phase de professionnalisation de la philosophie française après la seconde

guerre mondiale, qui se définit entre autre par une porosité plus grande vers d'autres disciplines dont l'histoire et les sciences sociales.

Enfin, l'on peut également à partir du cas Foucault reprendre la question de la créativité intellectuelle et de la transmission d'un style, ou de la manière de traiter les idées. Réflexion importante puisqu'il semble, encore, que l'originalité d'un produit textuel « philosophique » soit liée, en grande partie, à la méconnaissance de son processus de fabrication.

Bibliographie

- L'archive numérique Jean-Toussaint Desanti*. Lyon : ENS de Lyon, UMR 5037 Institut d'Histoire de la Pensée Classique, 2010. <<http://institutdesanti.ens-lyon.fr/spip.php?rubrique4>>.
- AT : *Archivists' Toolkit*. 2006–2009. <<http://archiviststoolkit.org/>>.
- Bellon, Guillaume. « Je crois au temps... » Daniel Defert, légataire des manuscrits de Michel Foucault, propos recueillis par Guillaume Bellon. » *Recto/Verso* 1 (2007) : 1–7. <<http://www.revuerectoverso.com/spip.php?article29>>.
- Bellour, Raymond. *Le livre des autres, essais et entretiens*. Paris : L'Herne, 1971.
- Bustarret, Claire. « Approche codicologique du manuscrit moderne. » *Critiques génétique, concepts, méthodes, outils*. Ed. Olga Anokhina et Sabine Petillon. Paris : IMEC Editeur, 2009. 47–60.
- Burnard, Lou. *Text Encoding for Interchange : A New Consortium*. 2000. <<http://www.ariadne.ac.uk/issue24/tei>>.
- EAD : *Description Archivistique Encodée. Dictionnaire de balises*. Trad. de l'anglais par le groupe AFNOR CG46/CN357/GE3. Society of American Archivists. (SAA), 2004. <<http://www.archivesdefrance.culture.gouv.fr/static/1066>>.
- Foucault, Michel. *Dits et écrits, T. IV*. Ed. Daniel Defert et François Ewald. Paris : Gallimard, 1995.
- Foucault, Michel, *Histoire de la folie, Extraits du dossier manuscrit préparatoire*. Portail Michel Foucault. Caen : IMEC, 2010. <<http://michel-foucault-archives.org/spip.php?article17>>.
- Foucault, Michel. *Les mots et les choses : une archéologie des sciences humaines*. Paris : Gallimard, 1990.
- Genette, Gérard. *Palimpsestes : La littérature au second degré*, Paris : Éditions du Seuil, 1982.
- Grafton, Anthony. *Les origines tragiques de l'érudition. Une histoire de la note en bas de page*. Paris : Éditions du Seuil, 1998.
- IMEC : *Institut Mémoires de l'Édition Contemporaine*, 2010. <<http://www.imec-archives.com/>>.
- ISAD(G) : *Norme générale et internationale de description archivistique : adoptée par le Comité sur les normes de description, Stockholm, Suède, 19–22 septembre 1999*. Ottawa : International council on archives (ICA), 2000. <<http://www.ica.org/fr/node/30001>>.
- Jacob, Christian. *Lieux de savoir, Vol. 1. Espaces et communautés*. Paris : Editions Albin Michel, 2007.
- Kraus, Dorothea. « Appropriation et pratiques de la lecture. Les fondements méthodologiques et théoriques de l'approche de l'histoire culturelle de Roger Chartier. » *Labyrinthe* 3 (1999). 13–25. <<http://labyrinthe.revues.org/index56.html>>.

- Latour, Bruno. « Les vues de l'esprit, une introduction à l'anthropologie des sciences et des techniques. » *Culture technique* (1985) : 4–30.
- Manuel d'encodage : Faire un répertoire ou un inventaire simple en EAD (Description archivistique encodée)* : Groupe d'experts AFNOR CG46/CN357/GE3. 2005–2009.
<<http://www.archivesdefrance.culture.gouv.fr/static/3322>>.
- MASTER : *Manuscript Access through Standards for Electronic Records. Reference Manual for the MASTER Document Type Definition – Discussion Draft*. Ed. Lou Burnard for the MASTER work Group. Oxford : Oxford University Computing Services, 2001.
<http://www.tei-c.org/About/Archive_new/Master/Reference/oldindex.html>.
- Müller, Bertrand. « Les lieux de savoir. Un entretien avec Christian Jacob. » *Genèses* 76 (2009) : 116–136.
- Neveu, Franck. *Dictionnaire des sciences du langage*. Paris : Armand Colin, 2004.
- Pleade*. Bordeaux : Société AJSLM, 2009–2010. <<http://pleade.com/>>.
- Rastier, François. *Arts et sciences du texte*. Paris : Presses Universitaires de France, 2001.
- Roques, Pascal. *UML. Modéliser une application web*. Paris : Eyrolles, 2008.
- Spiro, Lisa. *Archival Management Software. A Report for the Council on Library and Information Resources*. January 2009. <<http://www.clir.org/pubs/reports/spiro2009.html>>.
- TEI : *Text Encoding Initiative*. TEI Consortium, 2010. <<http://www.tei-c.org>>.
- Manuscript Description* : <<http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/MS.html>>.
- Willer, Mirna. *UNIMARC Manual*. München : K. G. Saur, 2007.

Putting the Text back into Context: A Codicological Approach to Manuscript Transcription

Elena Pierazzo, Peter A. Stokes

Abstract

Textual scholars have tended to produce editions which present the text without its manuscript context. Even though digital editions now often present single-witness editions with facsimiles of the manuscripts, nevertheless the text itself is still transcribed and represented as a linguistic object rather than a physical one. Indeed, this is explicitly stated as the theoretical basis for the *de facto* standard of markup for digital texts: the *Guidelines* of the Text Encoding Initiative (TEI). These explicitly treat texts as semantic units such as paragraphs, sentences, verses and so on, rather than physical elements such as pages, openings, or surfaces, and some scholars have argued that this is the only viable model for representing texts. In contrast, this chapter presents arguments for considering the document as a physical object in the markup of texts. The theoretical arguments of what constitutes a text are first reviewed, with emphasis on those used by the TEI and other theoreticians of digital markup. A series of cases is then given in which a document-centric approach may be desirable, with both modern and medieval examples. Finally a step forward in this direction is raised, namely the results of the Genetic Edition Working Group in the Manuscript Special Interest Group of the TEI: this includes a proposed standard for documentary markup, whereby aspects of codicology and *mise en page* can be included in digital editions, putting the text back into its manuscript context.

Zusammenfassung

Im Gegensatz zu früheren wissenschaftlichen Textausgaben bieten heute digitale Editionen von singular überlieferten Texte meist auch das Faksimile der Handschrift. Dennoch wird dabei der Text weiterhin vor allem als ein linguistisches und nicht als ein materielles Objekt transkribiert und präsentiert. In der Tat ist dies die explizit formulierte theoretische Grundlage der Richtlinien der Text Encoding Initiative (TEI), dem *de facto* Standard für die Auszeichnung digitaler Texte. Dieser Standard behandelt Texte als semantische Einheiten wie Paragraphen, Sätze, Verse usw., nicht jedoch als materielle Einheiten wie Seiten, Doppelseiten oder Oberflächen. Manche Philologen bezeichnen diese Herangehensweise sogar als einzig verlässliches Modell zur Repräsentation von Texten. Dem entgegen wird in diesem Beitrag argumentiert, die

Transkription von Text auf dem Dokument als einem materielles Objekt zu begründen. Hierzu werden zunächst die theoretischen Grundlagen des Textbegriffes betrachtet, wobei der Fokus auf dem Textbegriff der TEI und anderer Theorien des digitalen Markups liegt. Dann werden anhand von mittelalterlichen und modernen Beispielen eine Reihe von Gründen benannt, warum eine Herangehensweise, die das Dokument in den Mittelpunkt rückt, wünschenswert erscheint. Schließlich wird eine alternative Philosophie der Textauszeichnung diskutiert, die aus der Arbeitsgruppe “Genetische Edition” der “TEI Manuscript Special Interest Group” resultiert: ein Vorschlag zur Standardisierung dokumentarischer Textauszeichnung, die Aspekte der Kodikologie und *mise en page* in digitale Editionen integriert und damit den Text zurück in den Kontext seines materiellen Trägers bringt.

1. Introduction

In any branch of manuscript studies (editing, codicology, palaeography, art history, history) the first level of enquiry always is (or should be) the document, the physical support that lies in front of the scholar’s eyes. The fact that the text was transmitted to us by means of a specific physical object which has been organised in a certain way and preserved in one place or another has all sorts of consequences in the way we understand and receive that text. To understand the text that is contained in the manuscript, a deep study of the manuscript itself is fundamental: the layout, the type of script, the type of writing support, the binding and many other aspects can tell us about when, where, how and why this particular text was written in the page. It is also worth noting that the manuscript as object is increasingly becoming the object of study itself. This is based partly on the principle that a text cannot be understood outside its context, but also that the manuscript itself can tell us things that a text cannot, particularly if one is interested in the person or people who compiled it and the intellectual *milieu* in which it was compiled. On the other hand it is also very difficult to understand how, when and where a particular manuscript was produced without understanding the text(s) that it contains and the cultural circumstances that determined its production. Texts and documents live and make sense only within each other.

Nevertheless when it comes to transcribing and editing, the text is often taken out from its physical support, its context, and considered on its own, with only little, if any, evidence retained that it was once within a specific manuscript. This is the case, for instance, with the first edition of Jane Austen’s minor works, which were published by Chapman from the 1920s and collected in a single volume in 1954: from this edition (and all subsequent re-editions) the evidence that some of those texts come from heavily annotated draft manuscripts is missing, with the consequence that the texts and her

writing habits have been misunderstood by more than one reader.¹ The words used by Peter Schillingsburg about the difference in the implications and interpretation between print and digital can serve here as well:

Meanings are generated by readers who have learned to deal with symbols and formats. Change the symbol and the meaning changes; change the format and the implications are changed; change the contexts of interactions with texts and the importance and significance of the text changes.

(Schillingsburg 2006 146)

2. The TEI Guidelines and the Encoding of Documents

In the case of digital editions, this centrality of the text is encouraged by the structure and principles of the most prestigious standard for text encoding, the one produced and maintained by the Text Encoding Initiative. The approach of the TEI, in fact, forces scholars to consider the text first. The TEI certainly offers a very sophisticated way of describing manuscripts; however, when it comes to transcription, of the two main hierarchies (text and document) the TEI privileges the text, relegating topographical description to empty elements (<pb/>, <lb/>, <cb/>) or attributes (<add place="...">, <note place="...">); it is no coincidence, after all, that it is called the *Text* Encoding Initiative. The TEI does not say that documents are not relevant, but rather that they are less relevant than texts; to use a metaphor from bibliography, texts are “substantial” while documents are “accidental”.²

By using TEI, we have learnt to distinguish how to mark a text for what it really is (using descriptive markup) from what a text will look like when it will be output in print or on the screen (using procedural markup) and we have learned how this will help us in managing and preserving our data at best. If we are transcribing and encoding a text from a primary source (be that source manuscript or print) then we have also learned to use graphical features present in the source as a way to de-code the (ambiguous) code of that source. For example, if some string in the source document is in italic, we now wonder why is it so (following to the descriptive approach): is it perhaps a title, a foreign word, or for emphasis? Again, if something is written in the margins of a manuscript page, we wonder if it is an annotation, a variant, an addition: the fact that is in the margin or, say, in the interlinear space, does not change the nature of the text in this respect. All of the above can be done without considering “*accidental*” features

¹ One example is Virginia Woolf who, misled by the appearance of *The Watsons* in print, imagined Austen writing very bare sentences and then coming back to add the “flesh”; in contrast the evidence of the manuscripts suggests that the process worked the other way around, with the author “scratching out” superfluous words. See Sutherland 2005 140.

² This terminology of substantials and accidentals is in the sense established by Greg 1951.

(such as current lineation) or arbitrarily marked regions” (Renear 2004 223). In bringing about this approach, the TEI has

succeeded [...] [in] the development of a new data description language that substantially improves our ability to describe textual features, not just our ability to exchange descriptions based on current practice.

(Renear 2004 235)

This is valuable and important if you want to encode texts. However, as will be discussed shortly, there are many reasons why we might want to record the appearance of the source: that the string is in italics, for example, as well as or instead of why it is in italics. According to Renear (2004), this means that we are using procedural instead of descriptive markup. However, when we are trying to capture what the source document looks like, it is because we believe that this is at least as meaningful as the text it contains: we are documenting our source, not formatting our output, and so our encoding is descriptive, not procedural. In such a context, markup of pages, columns, lines, spacing, and so on may indeed be descriptive, not procedural (*pace* Renear 2004 224). As a matter of fact “what the text really is” depends on whether or not we think that Sperberg-McQueen’s fourth axiom (“[t]exts are linguistic objects”) is more, less or equally important than his fifth (“[t]exts occur in/are realized by physical objects”, 1991 37–40; see also below § 4).

TEI is based primarily on the principles of text-oriented markup, but it does make some significant concessions toward documentary markup by including elements like `<space/>` (“indicates the location of a significant space in the copy text”: Consortium 2009 § 11.6.1) or `<hi/>` (“marks a word or phrase as graphically distinct from the surrounding text, *for reasons concerning which no claim is made*”: Consortium 2009 § 3.3.2.2; our italics). The reason for providing such elements is that the scholar-encoder is not always able or willing to state why some textual features look the way they do.³ But while editors have the possibility of choosing between a semantically neuter `<hi/>` and the interpretative `<emph/>`, for instance, they cannot avoid the interpretational level when transcribing interlineated manuscripts, as they are only offered elements like `<add/>` or `<note/>`. Interestingly, the TEI also includes an element to capture page features: these are mainly for printed books and include `<fw/>`, which “contains a running head (e.g. a header, footer), catchword, or similar material appearing on the current page” (2009 § 11.7).⁴ But while the TEI offers a way to encode a header and footer, it does not provide

³ The Guidelines again: “If the encoder wishes to offer no interpretation of the feature underlying the use of highlighting in the source text, then the `hi` element may be used, which indicates only that the text so tagged was highlighted in some way. [...] The `hi` element is used to mark words or phrases which are highlighted in some way, but for which identification of the intended distinction is difficult, controversial, or impossible.” (2009 § 3.3.2.2). See also Sperberg-McQueen 1991 43–44.

⁴ That the element is intended for the printed page is clearly suggested by its full name: “forme word”.

a way to encode the pages which contain those headers and footers, only the *breaks* between pages.

In practice the elements mentioned above have proved insufficient for encoding texts within their physical context (as will be shown further below). As a result only two options have been available to scholars who wish to encode documents: either they have been convinced (or they have convinced themselves) that what they really wanted was to encode texts, perhaps also preserving some features of the original document but at a secondary level, or they have invented their own system to encode documents.

3. Why Documents

Before outlining a possible solution to these problems, it is necessary to understand why an editor might want to transcribe a text within its documentary context. Although not always recognised by the community, there are in fact very many such reasons. To list all of them is beyond the scope of this discussion; instead a necessarily short and somewhat arbitrary choice will now be presented.

3.1. The process to make the document is at least as important as the text

Scholars are not always interested in the text as a coherent flow of words: sometime they are interested in the process of production or in documenting how and why a given document was produced or a text composed. This is the case, for instance, in genetic criticism. Genetic criticism (or *critique génétique*) has characterised the French school of philology since the 1970s and is concentrated around the activities promoted by the ITEM (*L'Institut des textes et manuscrits modernes*). The theories and practices of genetic criticism have spread beyond France and are now considered to be fundamental scholarly approaches to the editing of any draft or working manuscript (*brouillons*). Compared to more traditional approaches to editing, genetic criticism privileges the analysis of the *process*, the stratified flow of authoring, as opposed to the “photograph” of the end result which is embodied by traditional diplomatic editions. This is one—but by no means the only—scholarly approach for which the study of the process is relevant, and any understanding of the process must surely begin with the document.

A recent facsimile and semi-diplomatic edition of a manuscript of the libretto of *Tosca* may exemplify this. A manuscript containing Puccini's working copy of the libretto of *Tosca* has been recently purchased by the Fondazione Cassa di Risparmio di Lucca which has encouraged and authorised an edition of this manuscript. While the text of *Tosca* is relatively well established and does not represent a problem in itself, it was known that Puccini was deeply involved in the composition of the libretto, together with Giulio Ricordi (a music publisher and a opera producer, to use the modern terminology) and the two *librettisti*, Luigi Illica and Giuseppe Giacosa. What was *not* known was the contribution of each of them and the way they used to work. The manuscript can give

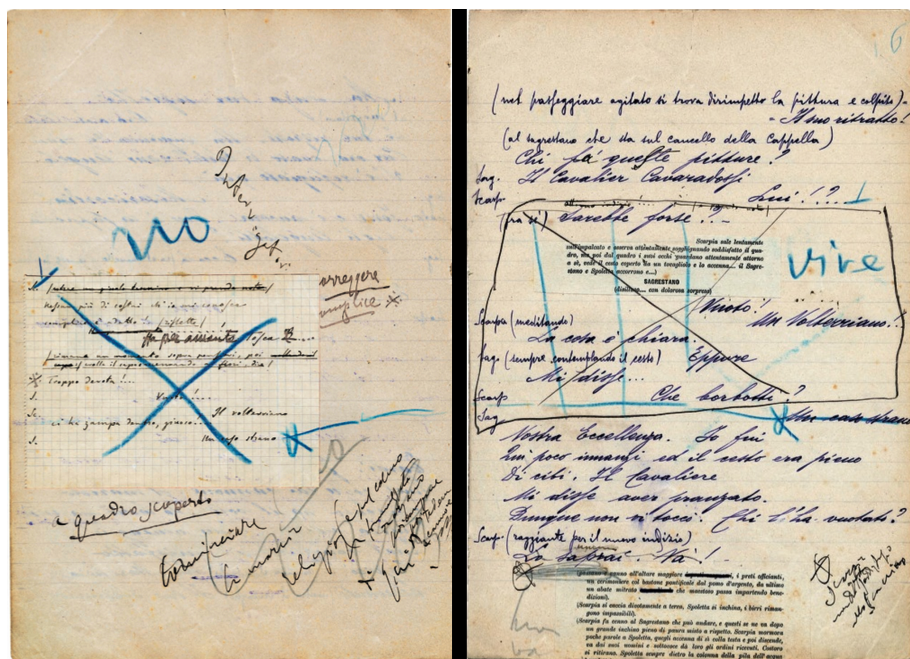


Figure 1. Tosca, pp. [38]–[39]. Reproduced by permission of Fondazione Cassa di Risparmio di Lucca.

an extraordinary insight into this matter, and the wonderful edition edited by Gabriella Biagi Ravenni is built around this principle (Sardou et al. 2009).

The working copy was most likely prepared within Ricordi's publishing house and consists of a stationery notebook, on the odd pages of which a version of the libretto has been transcribed by a professional scribe, interspersed with patches glued from an earlier printed version. The even pages were initially left blank for comments and annotations and now contain several layers of annotations by different hands which often spread onto the facing pages. The basic unit in the document is therefore the opening, the pair of facing pages as the book lies before the reader, and not the single page, as exemplified by Fig. 1.

In the transcription of the transmitted text the focus of Biagi Ravenni has been on the stratification of the hands (which she represents with different colors of ink), the temporal succession of the variants, and the disposition of words on the page. The semi-diplomatic edition reproduces the layout of the document, including the patches.

The result is not a text to be read in a traditional way, from the first to the last word, but instead each area tells a story and presents the intellectual effort of many people in producing the libretto. The document records the discussions, the thinking and the rethinking, the making and the unmaking of a process that lasted four years. The edition tries to make such a process explicit.

The fact that the edition of *Tosca* is not a text to be read should not surprise. According to Daniel Ferrer, “the draft is not a text, or a discourse, it is a protocol for making a text” (1998 261). This is demonstrated very clearly by the *Tosca* manuscript, and also by some of the pages of James Joyce which have been studied by Hans Walter Gabler. Gabler says:

Thus, when I look at—for example—two pages from James Joyce’s initial (disjunctive) draft for the “Circe” episode of *Ulysses*, my first concern is not to figure out what text the pages articulate. It is rather to find out how, as pages, they were successively filled. This means to analyse the patterns of ink and pencil on paper in terms of their inscriptional characteristics (that is: palaeographically), of their absolute positioning (that is: topographically), and of their relative positioning (that is: bibliographically).

(Gabler 2007 201)

Another example of the process being as important as the text is given by medieval *libri vitae*. These are manuscripts in which religious houses recorded the names of people associated with that house in order to pray for their souls. In general, the surviving manuscripts often contain a core of names which were written in a single block when the manuscript was first produced, and then many further names which were subsequently added, sometimes by hundreds of scribes over many hundreds of years. One example which has been recently edited comes from Durham: it is a very complex manuscript in which hundreds of hands have been identified (Rollason and Rollason 2007). By disentangling the scribal hands and dating the stints scholars can reconstruct the genesis of the manuscript and, perhaps more importantly, witness the growth and development of a community. In this case more than 1,100 additions have been counted, ranging from the mid 9th to the 16th century. A similar example is the *Winchester Liber Vitae* which has been published in a facsimile edition, the editor of which has written the following about just one page and its opening:

The present untidy appearance of the page is thus the end product of a highly complex process, representing the accumulation of names on many separate occasions in the eleventh and twelfth centuries. Yet the extraordinary sequence of entries on this opening [...] provides a striking reflection of the changing course of events at Winchester from the closing years of the Anglo-Danish dynasty, into the reigns of Edward the Confessor and Harold, and onwards past the Norman Conquest into the Anglo-Norman period and beyond.

(Keynes 2001; cf Keynes 1996 96–98)

Not only can we see the community grow and develop, here by studying the names and noting their different ethnicities at different dates, but we can see glimpses into particular historical events as well, such as royal visits to ecclesiastical institutions which show us not only which institutions the king supported but also who he was travelling with at the time (for examples of which see Bolton 2008 *passim*).

Despite the significance of the genesis and layout of these books, the printed editions have chosen largely to ignore this and to print the names simply as one long list, with minimal attention to layout, if any. The best-known edition is that of Jan Gerchow who included a large set of *libri vitae* but preserved none of the manuscript layout at all, printing the names in continuous lines across the page as if regular prose (1988 304–20). A subsequent edition of the core of another *liber vitae*, this one from Durham, attempted to preserve the layout but ignored the genesis by excluding all additions (Dumville and Stokes 2001); the layout of even these core names is perhaps significant, however, since they were written in alternating gold and silver, a feature which was carefully preserved in the edition. Indeed, one of the principle reasons for producing this edition was precisely because the layout was felt to be so important even in the core, and because it had not been preserved in the previous editions. Finally, the most recent complete edition of the Durham manuscript is exactly two thousand pages long, in three volumes and with a digital facsimile, but it still prints the text without any reference to *mise en page*, instead printing the names continuously but ordering them first by manuscript page and then by date of scribal stint (Rollason and Rollason 2007 vol. 1).

3.2. The text is determined by the document

The physical constraints of the page

In some cases the content of a document is determined by the document itself: for example, if the author has only a limited amount of paper, he or she will probably modify his or her authoring intention to fit the space available. This is often the case with correspondence, especially modern, informal correspondence. The correspondence

of Puccini gives some clear evidence of this, such as the post card sent to Albina Magi (Puccini's mother) on the 10th of November 1880 reproduced in Fig. 2.

Puccini was in Milan to attend the local Academy of Music and was always short of money; he was using this as an excuse not to write home; therefore his mother used to give him prepaid postal cards to encourage communication. The availability of space is limited here: Puccini first writes a sort of a "normal" letter to his mother, then when he runs out of space, he starts to write telegram-like, a-grammatical sentences in every available blank space which remains, not only in the margins but even between lines. We have no idea about the temporal-logical succession of such additions, but we can assume that he would have written fully developed sentences and paragraphs if he had more space. Transcribing the text as a single flow implies an arbitrary decision on behalf of the editor about the logical succession of the parts.

The phenomenon of the physical constraints of the page limiting the text is usually associated with modern manuscript materials. This is for good reason, since it is normally only in this context that draft versions of texts survive, and so only here can we see the composition taking place. There are certainly examples in medieval manuscripts of scribes adjusting their writing to fit into the page, either as additions being crammed into spaces on the page (examples of which are legion) or the main text being compressed or indeed extended to neatly fill the available space. Despite the very many examples of these two kinds, it is often unclear whether this adjustment has any impact on the text: unambiguous evidence of scribes omitting material from their exemplar for this reason is rare. In some cases, however, we can see the production process "in action", so to speak, and sometimes here we find strong hints, if not concrete evidence, of texts being constrained by the physical space. One such example is a rare case of correspondence from the early Middle Ages which perhaps survives in the form in which it was originally drafted. The document survives as *Canterbury, Dean and Chapter, C.1282*, and consists of a letter written by one Earl Ordlafe to King Edward "the Elder" some time in the period between AD 897 and 901. This document has been the subject of intense scholarly interest for a long time, and this interest has included close study of the language and phrasing, not least because it is a rare example of vernacular literacy from the lay nobility (Keynes 1990 248–9; Keynes 1992; Gretsche 1994; Hough 2000; Brooks 2009). However, as some scholars have noticed (but many have not), the text seems to end fairly abruptly, and furthermore the last line is crammed onto the bottom of the sheet of parchment; the evidence seems to suggest that Earl Ordlafe ended his letter in this way simply because he ran out of space on the page that he had available. Granted this interpretation is arguable, and has indeed been argued, but the point remains that the text here may be determined by the document, and so scholars who wish to understand the text must at least be aware of this possibility and must have the evidence at their disposal to evaluate the impact this may have on their own arguments.

Gatherings: can we understand the text without understanding the organisation of the document?

We have seen examples where the text cannot be fully understood without the document. There are other cases where the document and its codicology are required to understand the text, but also knowledge and understanding of the text is required to understand the document. One particularly well-known example of this is the so-called *Beowulf* manuscript. *Beowulf* is arguably one of the most important literary texts in English; it is an epic poem written in Old English, just over three thousands lines long, and has been the subject of almost innumerable scholarly and popular articles, books, translations, and adaptations, both written and cinematic. More recent scholarship has demonstrated that our understanding of this text is heavily determined by the codicological structure which preserves it. The only surviving copy of the poem is in a manuscript with several other works, and the relationship between these works has been debated extensively (Sisam 1953 61–96; Clement 1984; Kiernan 1996; Gerritsen 1998). However, as many scholars have failed to note, the quires of the manuscript were rearranged at some point, and there is even some evidence that *Beowulf* was once bound separately and that the manuscript as it survives today was originally conceived as two or even three separate volumes (Förster 1919 10–23 and 76; Ker 1957 281; Malone 1963 17 and 119; Clement 1984; Kiernan 1996 120–69). Similarly, another important debate relates to some damaged folios in the middle of the manuscript. The principle figures in this debate are Kevin Kiernan and Leonard Boyle: the former has suggested that this damage results from deliberate attempts by one of the scribes to erase and rewrite part of the text, whereas the latter has explained it by suggesting that the damaged pages were at a boundary between gatherings and that the gatherings were left unbound and were thereby exposed to water (Boyle 1981; Kiernan 1981; Kiernan 1996). Although this may seem like academic hair-splitting, the implications are very far-reaching, since Kiernan has used this position to argue repeatedly that the surviving manuscript represents an authorial copy of *Beowulf* and therefore that our understanding of the poem as composed in the eighth century or earlier is fundamentally wrong. He has also argued for revisions in editorial practice, since editors have tended to assume that our only surviving copy is a late and corrupt one and have therefore tended to intervene quite heavily in the text (Kiernan 1981; Kiernan 1995; Kiernan 1996 272–8 and *passim*). The implications of the codicology extend beyond *Beowulf*, too: the same manuscript also contains the only surviving copy of *Judith*, another important poem in Old English. This poem is written in the section of the manuscript which was certainly moved from its original position, and the text is now missing its beginning: how much is missing is unknown, but attempts to estimate the number of lines have been attempted based on the codicological evidence. These estimates have varied by orders of magnitude, and very different interpretations of the text have arisen as a result (Lucas 1990; Kiernan 1996 150–51).

These considerations are all very important for our understanding of these poems and therefore stand as examples of how our understanding of the text depends heavily on our understanding of the organisation of the document: in particular, discussion of these poems has been transformed by Kiernan's highly controversial interpretation of the them, but as Clement noted "[t]he collation [of the manuscript] is exceptionally important to Professor Kiernan's thesis" (Clement 1984 13). Unfortunately much of the evidence for this document's organisation and structure was destroyed when the manuscript was badly damaged by fire in 1731 (Prescott 1997); heated scholarly debate has ensued as a result, and at least six different and conflicting quire-structures have been published (Förster 1919 10–23; Dobbie 1953 xv–xvi; Ker 1957 282; Malone 1963 14–16; Boyle 1981; Kiernan 1996). For this reason, a detailed documentary edition which included full codicological evidence would be invaluable. In this case the structure of the gatherings themselves has been destroyed, as noted above, and so other forms of evidence must be preserved instead. Detailed measurements of the writing-frame, the exact distance between lines, on which side of the page the ruling was made, the arrangement of hair and flesh: all of these have been used as evidence for understanding the text, and all of them could usefully be encoded in a transcription of this manuscript. Indeed, it seems significant that Kevin Kiernan's own digital edition of *Beowulf* contained almost no codicological information in the transcription itself—all of this was relegated to the introduction or to his book, *Beowulf and the Beowulf Manuscript* (1996).

Another similar example is a pair of manuscripts which are now bound as one, along with fragments of a related third manuscript. These are all cartularies, that is, manuscripts containing documents which were issued originally as charters on single sheets of parchment but which were then copied into one book for administrative and organisational purposes. The three manuscripts in question were all produced at Worcester during the eleventh century: one, *Liber Wigornensis*, probably during the first or second decade of that century; the second, the "Nero-Middleton Cartulary", during the episcopate of St Wulfstan of Worcester (1062–1095); and the third, "Hemming's Cartulary", in the last decade of the eleventh century (Ker 1985; Tinti 2009 479). As Francesca Tinti has shown, the arrangement of the texts in the cartularies is significant and reveals much about the organisational and administrative practices in Worcester (Tinti 2002; Tinti 2009). *Liber Wigornensis* in particular is arranged in sections, and when the scribes finished a section they left the remaining pages in that gathering blank. In some cases these blank pages received further additions, but it also seems clear that the sections were rearranged at different times as the administrative principles changed (Tinti 2002; Baxter 2004 172–6; cf Tinti 2009 483–88). It is also worth emphasising that these issues are by no means limited only to these three manuscripts, but related questions also arise in other important sources for medieval history including Domesday Book, which is arguably the most important single source for the study of medieval

England and which also survives in two different forms, “Great” and “Little” Domesday: here even the spacing between words has proven significant in our understanding of this important pair of manuscripts (Galbraith 1961; Rumble 1985; Sawyer 1985 4). One author of this paper has been involved in discussions about a digital edition of the cartularies from Worcester, but these codicological issues leads to a number of complex scholarly requirements in any such edition. On the one hand, it is necessary to capture the current order of texts: this is a basic requirement of any edition of a single manuscript. In order to convey the different organisational principles, it is also necessary to capture the previous order (or orders) of the texts. This would then allow one to rearrange the material, presenting it in different ways according to the different arrangements. However, these prior arrangements are often difficult to establish, not least because the manuscript was damaged in the same fire as the *Beowulf* manuscript in 1731. For this reason, the editors of the proposed digital edition would like to allow scholars to rearrange the order of gatherings themselves, thereby allowing them to explore the material and test their own hypotheses. However, not all arrangements are equally likely or even possible. As with the *Beowulf* manuscript, codicological details such as ruling and the hair and flesh sides are all necessary to inform and constrain the possible arrangement of documents and quires. In this case the evidence and constraints are particularly complex, not least because they also depend on the arrangement of the text, and so the editors’ ideal may not be achievable in practice, but nevertheless the framework for encoding this information is still a desideratum.

Manuscripts of homilies often come in codicological units which have been rearranged at different times: again, our understanding of homiletic practice and the homilies themselves often depends on the arrangement of texts within the manuscript, and this in turn often depends on the codicology. Pamela Robinson has demonstrated that some medieval manuscripts, particularly homiliaries, once existed as separate booklets which were unbound and designed to be carried around for preaching (Robinson 1980; see also Rumble 1985 33–35, for the application of this to Domesday Book). Although these are now bound as single manuscripts, the evidence for their previous existence as booklets often survives, and if one accepts that a text is determined in part by its presentation and use (as argued by Schillingsburg, as discussed above) then it follows that this information is important. Again, if one wishes to understand the homilies as a collection—a topic that is often discussed in the literature (a necessarily small and arbitrary sample of which is given by Cross and Tunberg 1993; Clemoes 1966; Eliason and Clemoes 1966; Loyn 1971; Sauer 2000; Da Rold 2007; Treharne 2009)—then one must understand how this collection once functioned not as a single, fixed whole but rather as a set of distinct units which were designed to be rearranged at will.

Many other examples can be presented of manuscripts in which our understanding of the text depends on our understanding of the codicology (for another detailed example see Stokes forthcoming). In most cases they are similar to *Beowulf*, insofar as scholars

have recognised that many texts depend on their manuscript context, and that context depends in turn on the codicology. It is important to note that the emphasis here need not be on the process but can focus only on the result: we may not be concerned with the process by which the Beowulf manuscript came to be arranged the way it is today, but rather in understanding how it was arranged in the eleventh century. The process in itself is certainly an important research question, as we have already established, but, as these examples show, even the original structure is often important to understanding the text and can require detailed codicological information to be preserved in the encoding.

In addition to these examples where the text and its genesis is the subject of interest, it has already been noted that the manuscript as object is also very much a legitimate object of study (see Section 3, above); in some cases, however, the codicology cannot be understood without considering the text. To give one example, Oxford, Bodleian Library, Auct F.4.32 is an extremely complex manuscript which is built up of several different units which were written at different times and places and bound together in different stages. The relationship between these units is very difficult to establish, not least because it was rebound in the modern period, and it was presumably at this time when some bifolia were inserted the wrong way around.⁵ Even more complex is Cambridge, Corpus Christi College 367: M.R. James catalogued this manuscript as eight items in five distinct codicological units ranging from eleventh-century parchment to fifteenth-century paper (James 1912 II: 199–204), and again with folios misbound, reversed, different notes added at different stages, and so on (Stokes forthcoming). Our understanding of some of the texts in this manuscript depend utterly on their context: a booklist on folio 101v (*olim* 48v)⁶ can be dated and localised very closely because of the two texts that it stands between; similarly, our understanding of the Vision of Leofric, an account in Old English the only copy of which is preserved in that same section of the manuscript, changes significantly when we recognise that the scribe who wrote another text in that section is the “Hemming” of *Hemming’s Cartulary* (Stokes forthcoming; cf Baxter 2007 154–5 n. 6). How is one to present an edition of these manuscripts? If we present the text as it “should” be, with the folios put in their reading order, then we are not representing the manuscript. If we leave the folios as they are, then the text is unreadable. The obvious answer is: “This is a digital edition, we should present both the existing and the original arrangements.” The ideal digital edition would allow one to view each of the different units separately; to view the manuscripts as they were bound at different times; to view the manuscript as it is now, and as it is but

⁵ As well as Hunt’s facsimile edition (1961), the manuscript is now available online at ODL > Bodleian Library > MS. Auct. F. 4. 32.

⁶ I give here both the current foliation, established recently and apparently for the Parker on the Web project, and the previous foliation used in all published discussions to date, which restarts at the beginning of James’s Volume II (James 1912 II: 200). For discussion see Stokes forthcoming n. 3.

with the incorrectly bound pages back in order. All of this is possible, but only with a documentary view.

3.3. The text is graphically presented

With respect to the printed page, the manuscript page (especially, but not exclusively, in the modern era) is free of constraints and develops in many ways. We have already seen examples of this, such as the *Libri vitae*, and also Puccini's correspondence, where the written words more or less anarchically stratify on the writing space (Almuth Grésillon speak about a space where "la ligne horizontale y perd bien souvent ses droits", Grésillon 1994 51). In some other cases, we find manuscripts where the unconventional layout clearly reflects intentionality which is plastic or explicitly artistic, as in the famous *calligrammes* by Apollinaire⁷ or like the one by Jean Tardieu that is showing in Fig. 3.

Here the author is trying to represent with words the disposition of the mountains and hills that surround the Lake of Garda, and the reflection of those on the surface of the lake.⁸ Clearly a linearised transcription of the text will irremediably lose a fundamental part of the poem's meaning.

Examples in medieval manuscripts can also be found without much difficulty. The most striking examples are perhaps found in Islamic and Jewish manuscripts, such as the Hebrew micrography which seems to have developed in the tenth century.⁹ Although different in function to the modern ones, texts that are presented in graphical form are abundant in early Insular gospel-books, for example, such as the famous Lindisfarne Gospels or the Book of Kells. The opening page of each new book in the Lindisfarne Gospels is presented in a highly stylised and decorative format, so much so that the words are very difficult to read. This is illustrated by the "chi-rho" page in that manuscript, illustrated in Fig. 4, below. Many of these gospel books place special emphasis on this page, and scholars have suggested that intricate decorations like this were intended as something to be read like a text, to be meditated on and sought out in a nonlinear fashion as representation of the Godhead (Pulliam 2006 210; Brown 2003 77–8; cf the "Te igitur" pages as discussed by Suntrup 1980). Another of these gospel books is the Book of Armagh, which includes a page of readings from the Book of Revelations for which the scribe chose to arrange his text in a diamond format, illustrated in Fig. 5 below. Both the chi-rho page and the diamond-shaped one can be printed linearly, with abbreviations expanded and layout normalised, but a fundamental aspect of the page and its function is lost when this is done, and indeed it is significant that the second of the two manuscripts illustrated here was published in 1913 in an "*editio diplomatica*" which attempts to preserve the layout and some aspects of decoration

⁷ See some examples, for instance, in Apollinaire > Textes > Calligrammes.

⁸ Grésillon has provided the label of *écriture éclatée* for such type of writing (1994 57).

⁹ Many examples of these can be found online; for one starting-point see JTS ([n.d.]).

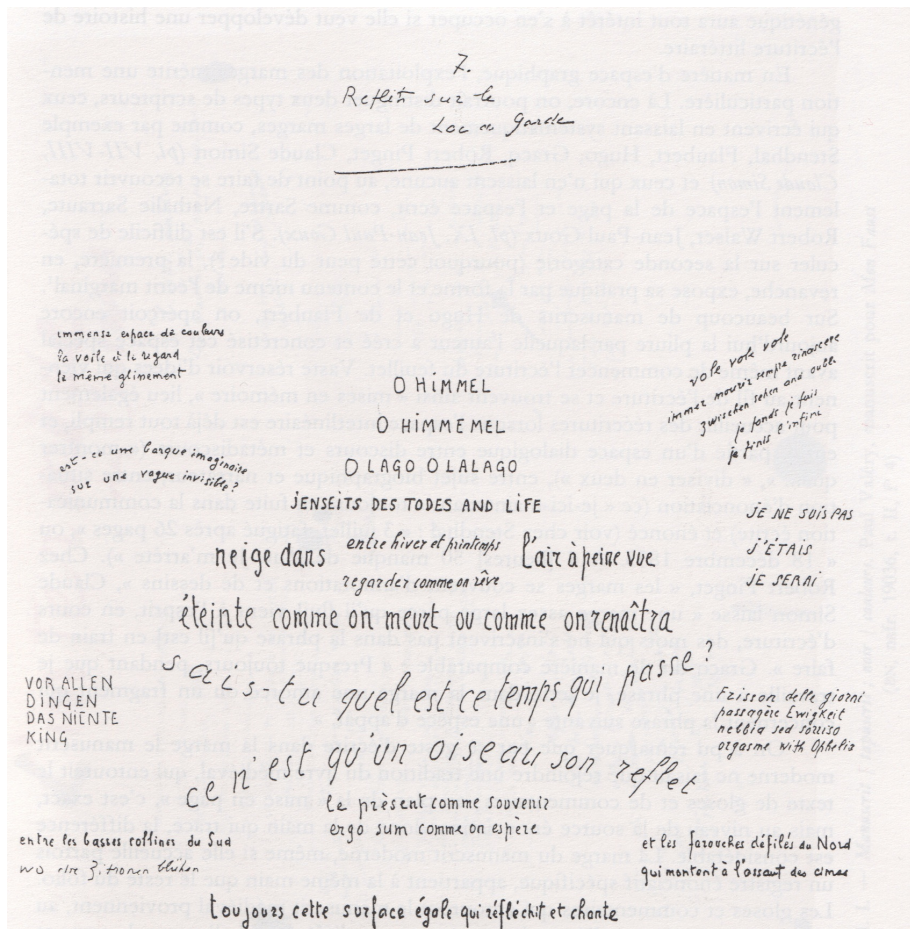


Figure 3. Jean Tardieu, *Reflet sur le lac de Garde* (reproduced from Grésillon 1994 59, which she reproduced from Tardieu 1990).

in the text (Gwynn 1913 i); a collotype facsimile of the first part of the book was also printed some years later (Gwynn 1937; for both publications see also *HyperStack*).

A further example is the set of so-called *carmina figurata* of Hrabanus Maurus, known as *De laudibus sanctae crucis* or sometimes *Opus in honorem sanctae crucis*. These survive as twenty-eight Latin poems, each of precisely thirty-six lines with thirty-six letters each which are designed to be laid out in a grid on the page; when so arranged, patterns are then formed by key letters which in turn spell out further words. An example is shown in Fig. 6, below.¹⁰ It is therefore a very early predecessor to the *calligramme* discussed above, and it brings much the same challenges. Unlike the *calligramme*, Maurus' text does retain some meaning when printed as a conventional text, but one must ask how much is lost in such situations, and even the *Patrologia Latina* edition of 1864 printed each poem twice, first in diagrammatic form and then as conventional verse immediately afterwards (Migne 1864 col. 141a–264d).

In addition to these examples where the texts are presented graphically, there are many other cases where the physical arrangement of the words on the page is critical to understanding the text. Maps are one such case, and several projects have produced or are producing digital editions of medieval *mappae mundi*, maps of the world.¹¹ Another example is the Bayeux Tapestry, which again contains image and text to narrate its story and which has also been published as a digital edition (Foys 2003). Although these may be seen as primarily diagrams, nevertheless they do contain text, sometimes in significant quantities, and this must be captured in any edition. However, even these are relatively straightforward compared to works like Peter of Poitiers' *Compendium historiae in genealogia Christi*, also known as the *Genealogy of Christ*, an extremely popular work which was first written at the end of the twelfth or start of the thirteenth century (Munroe 1978; Hilpert 1985). Although relatively short, usually filling about seven or eight manuscript pages, it contains a very large amount of information presented in a sophisticated layout which is both text and diagram, incorporating the two into one. The *Genealogy* is preserved sometimes as a manuscript, sometimes as a roll, and the content is presented with varying levels of sophistication and clarity; one of many examples is shown in Fig. 7, below, but photographs of numerous others can be found online.¹² It is hard to conceive of any meaningful edition of this work which

¹⁰ Further examples are Bologna, Collegio di Spagna 12 (reproduced at CIRSIFID-Irnerio), Lyon, Bibliothèque Municipale 597 (reproduced at BM-Lyon) and Österreichische Nationalbibliothek, Cod. Vind. 908, fol. 3v (reproduced in Sperberg-McQueen 1991 41, Fig. 3).

¹¹ Example projects which are complete or in progress include the *Digital Mappa Mundi*, by Martin Foys and Asa Mitman (2009), and the *Linguistic Geographies* project which focuses on the Gough Map (Kline 2001, Gough-Map 2010).

¹² At the time of writing, these include three entries in Digital Scriptorium (searching for “Compendium historiae”); the one preserved in Harvard is reproduced in full by Harvard University Library, Page Delivery Service (HUL-PDS). Another is preserved in Oxford, Bodleian Library, Lyell 71, 17v–28r which is reproduced at LUNA (searching for “Lyell 71”).



Figure 4. The Chi-Rho page of the Lindisfarne Gospels: London, British Library, Cotton Nero D.iv, 29r. Reproduced by permission of the British Library.

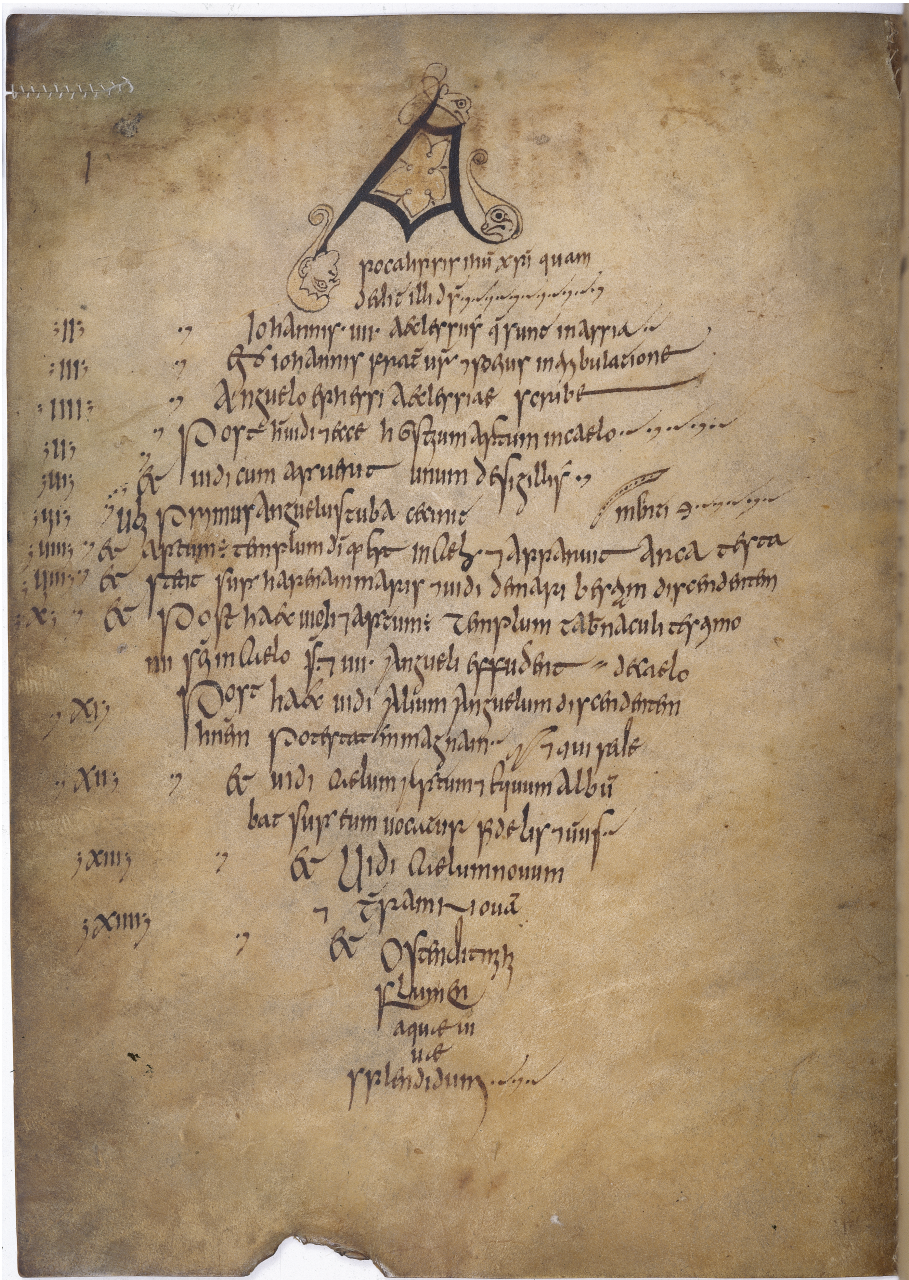


Figure 5. Readings from Revelations in the Book of Armagh: Dublin, Trinity College MS 52, 159v. Reproduced by permission of the Board of Trinity College Dublin.

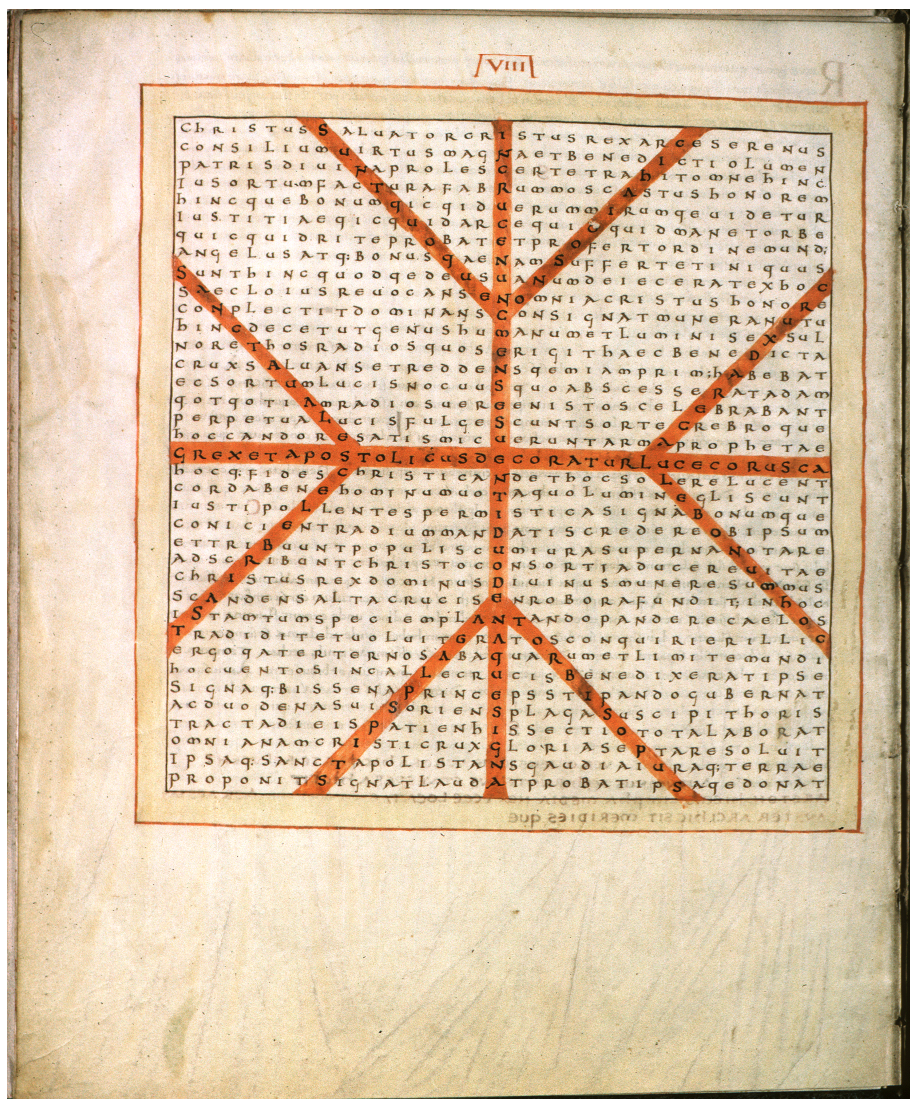


Figure 6. Hrabanus Maurus, *De laudibus sanctae crucis*. Bibliothèque municipale de Lyon 597, 5v. Photograph: Bibliothèque municipale de Lyon, Didier Nicole. Reproduced by permission of the Bibliothèque municipale de Lyon.

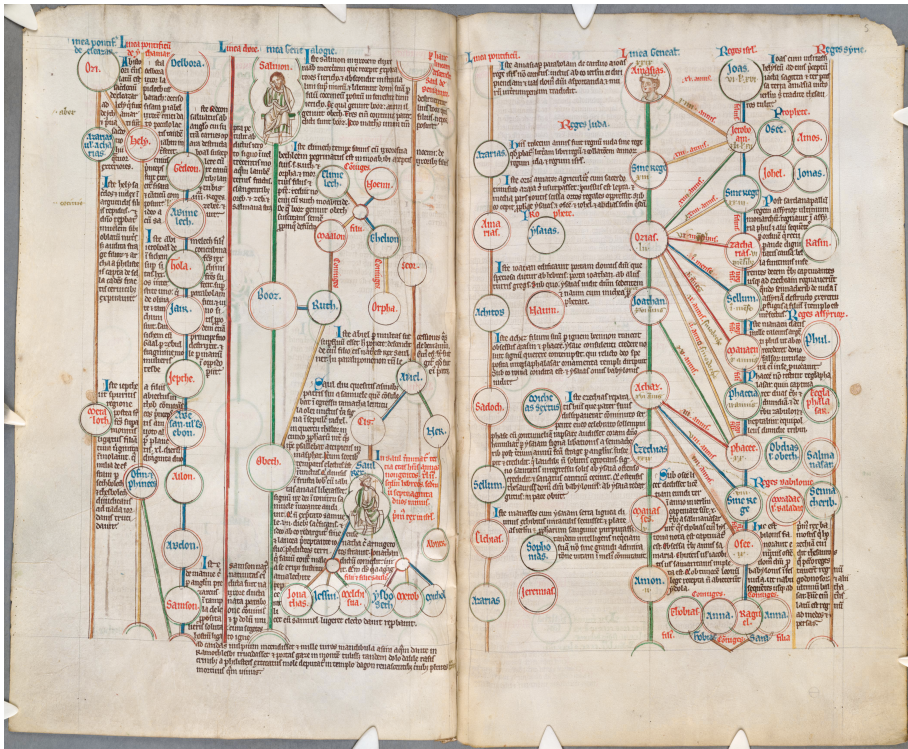


Figure 7. Peter of Poitiers, Genealogy of Christ. Cambridge, Corpus Christi College 83, 4v–5r. Reproduced by permission of the Master and Fellows, Corpus Christi College Cambridge.

does not somehow present it as both text and diagram, as neither is meaningful without the other.

Although arguably less extreme, another clear example of text which depends on layout for its meaning is the medieval gloss and commentary, in which a core text is given but with it is one or more layers of commentary. Many hundreds of manuscripts of this type survive and have been the subject of extensive study, with core texts including scripture or important and complex writers such as Boethius or Dante, or the *scholia* of Adam of Bremen. Even in the early medieval period, these manuscripts can acquire many layers of glossing, sometimes by tens of different scribes, in a complex system which includes various sets of interlinear and marginal additions. Initially these different layers glossed different aspects of the text: some might be linguistic, others

providing context, theological interpretation, and so on. In time, and particularly for biblical texts, medieval scholars started to produce commentaries of commentaries, and so the layering of glosses became more and more complex. From the twelfth century onwards, very complex and sophisticated page layouts were developed to accommodate these many interlocking texts (for examples see De Hamel 1987). The challenge, then, is how to edit these different texts while preserving the interconnections between them. Even simple linguistic glosses present problems, where (for example) the meaning of a Latin word is glossed with an alternative word written above it: as Raymond Page has reminded us (1992), more than one scholar has blundered due to editorial normalisations of these texts. If a “simple” case like this has led to scholarly error, then what of the very complex glossed bibles like that shown in Fig. 8, below?¹³ How can one accurately represent so many different texts and the relationships and connections between them without reproducing the layout of the page? In the past, some have attempted to print editions of these as simple, linearised texts (Meritt 1945; Meritt 1968), but this has resulted in significant loss of information at best, and disastrous blunders at worst.

3.4. There is no text

The final example to be considered here is draft manuscripts, in which the text is non-linear and can barely be defined text: as Daniel Ferrer reminds us, draft manuscripts are protocols, recipes to make a text (1998 261). Variations in the draft, also referred to as the *avant-text* or “pre-text”, have been explored and studied principally by the French school of genetic criticism. When a revision is present on the page, it means that the text existed in at least two versions, the one before and the one after the revision; the more variations that accumulate and stratify, the more versions of the same text can be deduced. In order to disentangle the paradigmatic variation, the different possibilities offered by the written page can be made explicit, such as in the way Almuth Grésillon has presented the genesis of a verse from the poem *Une étoile tire de l’arc* by Jules Supervielle (Grésillon 1994 165–67): here we count sixteen different versions, all implied and potentially contained by the stratified draft manuscript.¹⁴ In these cases we cannot speak of *the* text but of many possible texts, all enabled by the state of the document. A transcription should be able to offer the same possibility offered by the original manuscript, meaning that all possible readings should be present, not only the supposedly final will of the author.

¹³ For a similar example in print, see Sperberg-McQueen 1991 45, fig. 7.

¹⁴ For similar analyses of Giacomo Leopardi and Jane Austen, see Pierazzo 2009 182.

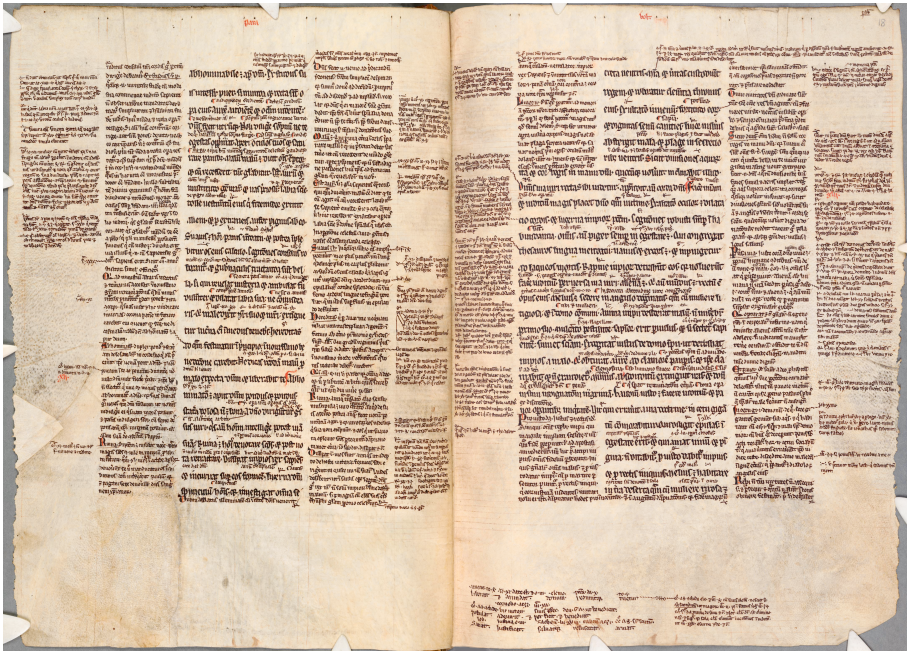


Figure 8. Glossed copy of Proverbs. Cambridge, Corpus Christi College 62, 17v–18r. Reproduced by permission of the Master and Fellows, Corpus Christi College Cambridge.

4. The OHCO View

When transcribing a text using the TEI encoding schema, one needs to take into account a few implicit theoretical assertions of “what text really is”, and the implications behind the adoption of a descriptive, non-procedural markup.

According to Renear (and others before him) descriptive markup has so many advantages that it must be right.¹⁵ More precisely, these scholars have argued that descriptive markup implies a single possible model of text—the only correct one—and that such a model postulates that texts consist of objects of a certain sort, structured in a certain way; in their view this means that texts have a linear order and they nest within each other without overlap. A text is then an “Ordered Hierarchy of Content Objects”, or OHCO (DeRose et al. 1990), a view which is shared by the TEI, even if

¹⁵ This position has been summarised by Renear 2004 225, citing his and others’ work, particularly DeRose et al. 1990.

only implicitly.¹⁶ Renear et al. derive from this statement that “A *book* for instance is a sequence of chapters, each of which is a sequence of major sections, each of which in turn is a sequence of subsections.” (Renear et al. 1996, § *OHCO-1*); it is perhaps worth noticing that here in order to exemplify “what a *text* really is”, the author has chosen to describe the structure of a *book*, thereby slipping from the immaterial abstraction of the text to the physical level of the document. However, as we speak about a physical entity (the book), we can also describe it as an object which is composed by a binding, a cover, and a sequence of pages which may or may not be organised in quires. This vision was not unknown to the OHCO working group, but it was rejected as not suited to scholarly purposes:

[a] book can be divided into pages; a page into the header, the main text area (perhaps with several columns, embedded pictures, etc.), an optional footnote area, and a footer. However, even this model fails to provide the kind of text handling needed by authors and scholars. How can one find equations, poetry quotations, lines of verse, and the like?

(DeRose et al. 1990 10)

But what if I don’t want to find poetry quotations? What if the purpose of my research “is rather to find out how, as pages, they were successively filled”, to use Gabler’s words (2007 201)? Is this not also a legitimate scholarly purpose?

In our earlier discussion we have demonstrated that, when considering texts that are contained within a manuscript, in order to say “what a text really is”, one must deal with the physical embodiment of that text.¹⁷ In our opinion the OHCO view represents a highly idealised and simplified vision of the text which does not take into consideration the modalities and the circumstances of the transmission of that text. In the real world, there are fundamental layers of interpretation that are missed when the text is taken out of its context. To use a couple of metaphors (with religious connotations), text and document are like Ying and Yang, or body and soul: neither can live without the other. While these metaphors may perhaps go too far, in that we can (and indeed often do) study the document without the text or vice versa, nevertheless we inevitably lose an integral part of the whole when we do so, and this will necessarily limit our study.

¹⁶ “The Guidelines of the Text Encoding Initiative exhibit a characteristically ambiguous stance: although they seem to privilege this view and benefit from its influence, they do not specifically invoke, explain, or defend it.” (Renear et al. 1996, Introduction)

¹⁷ Similar considerations can, of course, be applied to texts contained in printed books, but the case of manuscripts is perhaps even more evident for the reasons discussed above.

5. What to Do

How then can we encode both the textual and the documentary views, and how do these two views relate to each other? When we model an object to be studied we need to make sure we are not building a model that is as complex as the object of our study: a certain level of simplification and abstraction is required for the model to be useful. Decades of text encoding have shown that as soon as we try to mix texts and documents we encounter overlapping hierarchies: textual boundaries do not coincide with documentary ones, apart from some very specific cases, and to handle both at the same time is not possible, especially with a strictly hierarchical markup language like XML.

These two perspectives—textual and documentary—are therefore probably mutually exclusive in practice, but there is no reason why the former should always prevail over the latter: the choice between them should depend on the point of view of the researcher, on the nature of the document and on the intended use of the encoded material and not on the limitations of an given encoding schema. Different scholarly approaches are now paying growing attention to the physical object: to these scholars the contained text is, if not less important, certainly strictly dependant on the object in which it is preserved. These include the *Critique Génétique*, the *New Philology*, and the *Textual (or Analytic) Bibliography*, to mention just a few of them. Scholars who use these approaches should be offered a way to encode what they really want to encode in a standard way. At present the solution is either to invent a new encoding system from scratch or to heavily modify the TEI in order to fit the needs of the particular project, with each scholar and project doing so in a different way. In the following quotations Aurèle Crasson and Jean-Daniel Fekete discuss the need for the former, while Matt Cohen, director of the Interface Development for Static Multimedia Documents project at Duke University, for the latter.

TEI fonctionne bien pour les manuscrits relativement propres où le texte est stable mais ne convient plus lorsque les phénomènes paratextuels prolifèrent, comme c'est le cas dans les manuscrits littéraires modernes ou dans des brouillons.

(Crasson and Fekete 2004 168)

We [...] decided to table <facsimile> and extend the existing Whitman Archive schema such that it could handle pages—surfaces—as intellectually significant structural units.

(Cohen et al. 2007 2)

5.1. Encoding documents

Recently, the TEI has started to consider the possibility of encoding documents and not only texts, as we shall see shortly. For the TEI this represents a drastic new development away from the usual categories of textual analysis which have been consolidated over more than twenty years of experience with text encoding. As discussed above, TEI has been successful in the “development of a new data description language that substantially *improves* our ability to describe textual features” (Renear 2004 235); now the community needs the TEI to do the same for documents. For instance: what are documents made of? We know that, according to the TEI, texts are made fundamentally of structural divisions (chapters, sub-sections, poems, acts, scenes), and these do not contain text but further structural features (paragraphs, lists, tables, lines of poetry, speeches). But we do not yet know what documents are made of: quires? Pages? Double pages? Folios? Bifolios? Patches? And what do those units contain? Areas, regions or text?

The previous examples show how the encoding of documents should be able to address fundamental codicological questions on how the manuscript was presumably originally organised, to describe its present state and, possibly, to describe how we can go from one to another. In other words, the codicological encoding needs to be addressed not only from a descriptive point of view, but in time: it also needs to be genetic. The same examples show that the fundamental unit of transcription is not always or necessarily the page, but can be the opening, the bifolium or any surface that, according to the editor, represents the smallest meaningful subdivision of the physical object. In practice this can be almost anything. For example, the draft poem “America to Old-World Bards” was written by its author, Walt Whitman, on the back of old envelopes and letters, some of them glued together to form a bigger writing surface. In this case, everything is problematic, including the choice of what to transcribe and what not to: it is in fact worth noting that the editors have chosen to transcribe only those parts of the document that contain the poem, and not (for example) the front of the envelopes, the content of which is nevertheless used for dating the composition of the poem (Whitman 2005–2010).

How do we encode documents? Shall we just reverse the hierarchy text/document of the TEI encoding schema, regarding the document as “substantial” and semantic, textual markup as “accidental”?¹⁸ According to Crasson et al., “un seul niveau de description ne suffit pas pour capter la structure d’un manuscrit” (2004 168), meaning that if you chose to encode either the text instead of the document or the document instead of the text you will lose some layers of meaning contained in the original object. But, as noted above, perhaps the attempt to encode the text within the document is too ambitious and can lead either to ungovernable markup or to unreconcilable overlapping hierarchies. It

¹⁸ For a similar proposal see Pierazzo 2009 174–76.

seems that unless one level prevails, whether textual or documentary, the two levels cannot live together. This is because a scholar who is encoding page by page and line by line may wish to mark up textual features at both block level and at in-line level (examples of the former are paragraphs, stanzas, and speeches; the latter includes dates, names of people, and so on). However, these features can potentially overlap—and in practice they almost always will. From these early days in digital documentary transcription, it seems that a parallel encoding (texts and documents) is the way to go.¹⁹

5.2. The Proposal of the Genetic Edition Working Group

A proposal for adding a documentary view to the TEI has recently been accepted in principle by the TEI Council (April 2010): this means that the details of the elements and attributes may still be adjusted, but that the overall concept and theoretical basis has been agreed. This is part of a bigger proposal for the encoding of genetic editions which has been put together by a task force within the Manuscripts Special Interest Group.²⁰ The working group has recognised that it was impossible to deal with genetic criticism and modern manuscripts without first addressing the lack of support for encoding documentary features. The proposal has then been articulated in three main parts:

1. The documentary view (for which see below).
2. Transcription enhancement, which includes a set of new elements for encoding textual and para-textual features typical of working manuscripts. It includes, for instance, elements for re-writing or for functional annotation such as “move the paragraph here”. It also includes a generic element able to encode any type of phenomena without implying any one particular interpretation. For instance, when an editor sees that a word has been struck through in a given document, that editor can say either that the word has been deleted (encoding it at the interpretational level) or that there is a line on top of it (encoding at the documentary level).
3. Genetic markup, which includes a group of elements for encoding evolution across time and across the different manuscripts that a work or a document has had, going from its first documented elaboration to the “finished” product which is usually the published book or the manuscript in its current state.

The documentary view allows one to transcribe texts from a documentary perspective alongside or as an alternative to the textual perspective. According to this proposal, a

¹⁹ This is the choice discussed by Crasson and Fekete 2004 while presenting the *Transcripteur*, an open source editor designed to help in the encoding of modern draft manuscripts. This tool allows one to transcribe the text from two different points of view: documentary/diplomatic and textual; it also integrates images and gives the possibility of connecting the transcription to the facsimile.

²⁰ The Task force is chaired by Fotis Iannidis. Other members of the working group are Elena Pierazzo, Malte Rehbein and Lou Burnard, and fundamental contributions have also been made by Gregor Middell, Moritz Wissenbach and Paolo D'Iorio. See Pierazzo et al. 2010.

document can contain surfaces and surfaces can contain zones or patches (i.e. pieces of paper attached on top of the main surface). Zones can contain text, lines of text, or more zones. The terminology is deliberately generic, so “surface” could refer to a page, an opening, a face of a membrane or the side of a tapestry, according to the specific circumstance; the same applies to “zones” which could be marginal areas or any other polygonal area within a surface.

At present a dedicated way to encode codicological structure and its evolution is missing; it is nevertheless possible to use one of the new genetic structures or perhaps the generic TEI linking mechanism to group different surfaces together (for which see Consortium 2009 §16.1): the community of users has been called to use the new encoding and to create case studies and examples according to different scholarly needs. If the tools that have been offered prove inadequate then they will be encouraged to follow the example of the genetic editions working group and to propose an improvement to the TEI.

6. Conclusions

In the first volume of *Codicology and Palaeography in the Digital Age* and the associated conference, it was noted that much research on “digital” manuscript studies has been on palaeography, with relatively little attention paid to codicology (Stinson 2009 36). However, as the editors of the volume noted then, and as we hope this article has shown, codicology is crucial to the understanding of very many texts, particularly if page layout is included as part of this topic. The emphasis of most models for XML encoding, and especially the TEI, has been on modelling texts, and one may well argue that the TEI’s role should be this and no more: it is, after all, the *Text Encoding Initiative*. Nevertheless, very many texts cannot be understood as “pure” text without regard to context and layout, as linear sequences of tokens (Caton 2009 80) or even as ordered hierarchical content objects (DeRose et al. 1990).

The constraints of print technology have often relegated codicology to the introductions of text editions, where editors traditionally provide a more or less superficial description of the structure of source documents. The advent of digital publication has allowed for much less constrained types of edition: texts can be presented in multiple views (diplomatic, reading, glossed, etc.), facsimiles of the source documents are much more affordable, and in general the structure of publications is much more flexible. In the same ways as genetic editions (Pierazzo 2009 171–72), these new possibilities could be used to apply codicological methods and analysis in a much more effective way to digital editions, making them the engine able to drive a scholarly interpretation which is firmly aware of the implications and consequences of text transmission.²¹

²¹ Any co-authored paper normally involves close collaboration and this is no exception. Nevertheless, we shall attempt to delineate our respective contributions. Elena Pierazzo wrote sections 1, 2, 3.1, 3.4, 4, 5 and

Bibliography

- Apollinaire. *Guillaume Apollinaire*. Paris: ONAC, 2007–2009.
<http://www.guillaume-apollinaire.fr/>.
- Austen, Jane. *Minor works now first collected and ed. from the manuscripts by R. W. Chapman, with illus. from contemporary sources*. Ed. Robert William Chapman. London: Oxford University Press, 1954.
- Baxter, Stephen David. “Archbishop Wulfstan and the Administration of God’s Property.” *Wulfstan, Archbishop of York: the proceedings of the second Alcuin Conference*. Ed. Matt Townend. Turnhout: Brepols, 2004. 161–205.
- Baxter, Stephen David. *The earls of Mercia: lordship and power in late Anglo-Saxon England*. London: Oxford University Press, 2007.
- BM-Lyon. *Bibliothèque municipale de Lyon, Manuscrits Mérovingiens et Carolingiens*. Lyon: Ville de Lyon. <http://florus.bm-lyon.fr/>.
- Bolton, Timothy. *The empire of Cnut the Great: conquest and the consolidation of power in Northern Europe in the early eleventh century*. Leiden: Brill, 2008.
- Boyle, Leonard E. “The Nowell Codex and the Poem of Beowulf.” *The Dating of Beowulf*. Ed. Colin Chase. Toronto: University of Toronto Press, 1981. 23–32.
- Brooks, Nicholas. “The Fonthill Letter, Ealdorman Ordlafe and Anglo-Saxon Law in Practice.” *Early medieval studies in memory of Patrick Wormald*. Eds. Stephen Baxter et al. Farnham, Aldershot: Ashgate, 2009. 301–18.
- Brown, Michelle P. *The Lindisfarne Gospels: Society, Spirituality, and the Scribe*. London: British Library, 2003.
- Caton, Paul. “Lost in Transcription: Types, Tokens, and Modality in Document Representation.” *Digital Humanities 2009: Conference Abstracts*. Maryland (MD): Maryland Institute for Technology in the Humanities, 2009. 80–82.
- CIRSFID-Irnerio. *CIRSFID Progetto Irnerio*. Bologna: University of Bologna, 2003–2007.
<http://irnerio.cirsfid.unibo.it/>.
- Clement, Richard W. “Codicological Considerations in the Beowulf Manuscript.” *Essays in Medieval Studies*, 1 (1984): 13–27.
- Clemoes, Peter. “History of the Manuscript: Origin and Contemporary Correction and Revision.” In *Ælfric’s First Series of Catholic Homilies: British Museum, Royal 7 C. XII, fols. 4–218*. Eds. Norman Ellsworth Eliason and Peter Clemoes. Copenhagen: Rosenkilde and Bagger, 1966. 28–35 and 24–5.
- Cohen, Matt, Erika Fretwell and Kevin Webb. *White Paper: Interface Development for Static Multimedia Documents*. 2007. http://asmodeus.ws/cohenlab/NEH_White_Paper.pdf.
- Crasson, Aurèle and Jean-Daniel Fekete. “Structuration des manuscrits: Du corpus à la région.” *Proceedings of CIFED 2004*. La Rochelle (France), 2004: 162–168.
<http://www.lri.fr/~fekete/ps/CrassonFeketeCifed04-final.pdf>.
- Crasson, Aurèle and Jean-Daniel Fekete. *Transcripteur*. 2008
<http://tei-eclipse.gforge.inria.fr/transcripteur/>

the modern examples in sections 3.1.1, 3.2.1 and 3.3; Peter Stokes wrote sections 3.1.2 and 3.2.2 and the medieval examples in sections 3.1.1, 3.2.1 and 3.3.

- Cross, James E. and Jennifer Morrish Tunberg. *The Copenhagen Wulfstan Collection: Copenhagen, Kongelige Bibliotek, Gl. Kgl. Sam. 1595*. Copenhagen: Rosenkilde and Bagger, 1993.
- Da Rold, Orietta. "Homilies and Lives of Saints: Cambridge, University Library, li. 1. 33." *The Production and Use of English Manuscripts 1060 to 1220*. University of Leeds and University of Leicester, 2010. <<http://www.le.ac.uk/english/em1060to1220/mss/CUL.li.1.33.htm>>.
- De Hamel, Christopher. *Glossed Books of the Bible and the Origins of the Paris Booktrade*. Woodbridge: D.S. Brewer, 1987.
- DeRose, Stephen J., David G. Durand, Elli Mylonas and Allen H. Renear. "What is Text, Really." *Journal of Computing in Higher Education*, 1 (1990): 3–26. Online: <<http://delivery.acm.org/10.1145/270000/264843/p1-derose.pdf>>.
- Dobbie, Elliot Van Kirk, ed. *Beowulf and Judith*. New York: Columbia University Press, 1953.
- DS. *Digital Scriptorium*. Libraries Digital Program Division. New York: Columbia University Libraries. <<http://scriptorium.columbia.edu/>>.
- Dumville, David N. and Peter Anthony Stokes, eds. *Liber Vitae Dunelmensis*. Trial Version, 2 parts. Cambridge: Dept. ASNC, 2001.
- Eliason, Norman Ellsworth and Peter Clemoes, eds. *Ælfric's First Series of Catholic Homilies: British Museum, Royal 7 C. XII, fols. 4–218*. Copenhagen: Rosenkilde and Bagger, 1966.
- Ferrer, Daniel. "The Open Space of the Draft Page: James Joyce and Modern Manuscripts." *The Iconic Page in Manuscript, Print and Digital Culture*. Ed. Georges Bornstein and Theresa Tinkle. Ann Arbor (MI): University of Michigan Press, 1998. 249–267. <<http://www.item.ens.fr/index.php?id=23616>>
- Förster, Max. *Die Beowulf-Handschrift*. Berichte über die Verhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig, Philologisch-historische Klasse 71 (1919).
- Foys, Martin K., ed. *The Bayeux Tapestry: Digital Edition*. Leicester: Scholarly Digital Editions, 2003.
- Foys, Martin K. "Digital Mappaemundi: A Resource for the Study of Medieval Maps and Geographic Texts." *Library of Funded Projects*. National Endowment for the Humanities, Office of Digital Humanities, 2009. <<http://www.neh.gov/ODH/Default.aspx?tabid=111&id=9>>.
- Gabler, Hans Walter. "The Primacy of the Document in Editing." *Ecdotica*, 4 (2007): 197–207.
- Galbraith, Vivian Hunter. *The making of Domesday book*. Oxford: Clarendon Press, 1961.
- Gerchow, Jan, ed. *Die Gedenküberlieferung der Angelsachsen mit einem Katalog der libri vitae und Necrologien*. Berlin: Walter de Gruyter, 1988.
- Gerritsen, Johan. "The Copenhagen Wulfstan Manuscript: A Codicological Study." *English Studies*, 79 (1998): 501–511.
- Gough-Map. *Linguistic Geographies: The Gough Map of Great Britain*. Belfast: Queen's University, 2010. <<http://www.goughmap.org/>>.
- Greg, Walter Wilson. "The Rationale of Copy-Text." *Studies in Bibliography*, 3 (1950–51): 19–36.
- Grésillon, Almuth. *Éléments de critique génétique: lire les manuscrits modernes*. Paris: Presses universitaires de France, 1994.
- Gretsch, Mechthild. "The language of the 'Fonthill Letter'". *Anglo-Saxon England*, 23 (1994): 57–102.
- Gwynn, John, ed. *Liber Ardmachanus: The Book of Armagh*. Dublin: Hodges, Figgis & Co., 1913.

- Gwynn, Edward, ed. *Book of Armagh: The Patrician Documents*. Dublin: Stationery Office, 1937.
- Hilpert, Hans-Eberhard. "Geistliche Bildung und Laienbildung. Zur Überlieferung der Schulschrift *Compendium Historiae in Genealogia Christi* (*Compendium veteris testamenti*) des Petrus von Poitiers (1205) in England." *Journal of Medieval History*, 11 (1985): 315–31.
- Hough, Carole. "Cattle-Tracking in the Fonthill Letter." *English Historical Review*, 115 (2000): 864–892.
- HUL-PDS [Harvard University Library Page Delivery Service]. *Peter, of Poitiers, ca. 1130–1205. Compendium historiae in genealogia Christi: manuscript, [ca. 1200–ca. 1250]. MS Typ 216. Houghton Library, Harvard University, Cambridge, Mass.* Cambridge (MA): Harvard University Library. <<http://pds.lib.harvard.edu/pds/view/12438364>>.
- Hunt, Richard W., ed. *Saint Dunstan's Classbook from Glastonbury: Codex Biblioth. Bodleianae Oxon. Auct. F. 4.32*. Amsterdam: North-Holland Publishing Co., 1961.
- HyperStack: *Saint Patrick's Confessio Hypertext Stack Project*. Dublin: Royal Irish Academy, 2008–2010. <<http://www.confessio.dmlcs.org/>>.
- ITEM: *Institut des Textes et Manuscrits Modernes*. Paris: CNRS (Centre National de Recherche Scientifique) / ENS (Ecole Normale Supérieure) – UMR 8132. <<http://www.item.ens.fr/>>.
- James, Montague Rhodes. "The Earliest Inventory of Corpus Christi College." *Proceedings of the Cambridge Antiquaries Society*, 14 (1912): 89–104.
- JTS: *Micrography. The Hebrew Word as Art*. Library of the Jewish Theological Seminary, [n.d.]. <<http://www.jtsa.edu/prebuilt/exhib/microg/index.shtml>>.
- Ker, Neil R. *Catalogue of Manuscripts containing Anglo-Saxon*. Oxford: Clarendon Press, 1957.
- Ker, Neil R. "Hemming's Cartulary: A Description of the Two Worcester Cartularies in Cotton Tiberius A.xiii." *Books, Collectors and Libraries: Studies in the Medieval Heritage*. Ed. Andrew G Watson. London: Hambledon, 1985. 31–59.
- Keynes, Simon. "Royal Government and the Written Word in Late Anglo-Saxon England." *The Uses of Literacy in Early Mediaeval Europe*. Ed. Rosamond McKitterick. Cambridge: Cambridge University Press, 1990. 226–57.
- Keynes, Simon. "The Fonthill Letter." *Words, Texts and Manuscripts: Studies in Anglo-Saxon Culture Presented to Helmut Gneuss on the Occasion of his Sixty-Fifth Birthday*. Ed. Michael Korhammer. Woodbridge: D.S. Brewer, 1992. 53–97.
- Keynes, Simon, ed. *The Liber Vitae of The New Minster and Hyde Abbey Winchester: British Library Stowe 944: together with leaves from British Library Cotton Vespasian A. VIII and British Library Cotton Titus D. XXVII*. Copenhagen: Rosenkilde and Bagger, 1996.
- Keynes, Simon. *Images of Anglo-Saxon Manuscripts 5: The Liber Vitae of the New Minster, Winchester*. Cambridge: Trinity College. <<http://www.trin.cam.ac.uk/sdk13/LibVitNM.html>>.
- Kiernan, Kevin S. "The Eleventh Century Origin of Beowulf and the Beowulf Manuscript." *The Dating of Beowulf*. Ed. Colin Chase. Toronto: Toronto University Press, 1981. 9–21.
- Kiernan, Kevin S. "The Legacy of Wiglaf: Saving a Wounded Beowulf." *Beowulf: Basic Readings*. Ed. Peter S. Baker. New York: Garland, 1995. 195–218.
- Kiernan, Kevin S. *Beowulf and the Beowulf Manuscript*. Ann Arbor (MI): University of Michigan Press, 1996.
- Kline, Naomi Reed. *A wheel of memory: the Hereford Mappamundi*. Ann Arbor (MI): University of Michigan Press, 2001.

- Loyn, Henry. *A Wulfstan Manuscript, containing Institutes, Laws and Homilies: British Museum Cotton Nero A.I.* Copenhagen: Rosenkilde and Bagger, 1971.
- Lucas, Peter J. "The Place of Judith in the Beowulf-Manuscript." *Review of English Studies*, 41 (1990): 463–78.
- LUNA. Oxford: Bodleian Library, 2008–2010. <<http://bodley30.bodley.ox.ac.uk:8180/luna/servlet>>.
- Malone, Kemp. *The Nowell Codex: British Museum Cotton Vitellius A. XV, Second MS.* Copenhagen: Rosenkilde and Bagger, 1963.
- Marchetti, Arnaldo and Vittorio Giuliani, eds. *Puccini com'era*. Milano: Curci, 1973.
- Meritt, Herbert Dean, ed. *Old English Glosses*. London: Oxford University Press, 1945.
- Meritt, Herbert Dean. *Some of the Hardest Glosses in Old English*. Stanford (CA): Stanford University Press, 1968.
- Migne, Jacques Paul, ed. B. *Rabani Mauri. Fuldensis abbatis et moguntini archiepiscopi opera omnia*, Vol. 1. Patrologia Latina 107 (Paris, 1864).
- Munroe, William H. "A Roll-Manuscript of Peter of Poitiers' Compendium." *The Bulletin of the Cleveland Museum of Art*, 65 (1978): 92–107.
- ODL: *Early Manuscripts at Oxford University*. Oxford: Oxford University, 2000–2001. <<http://image.ox.ac.uk/>>.
- Page, Raymond I. "On the Feasibility of a Corpus of Anglo-Saxon Glosses: The View from the Library." *Anglo-Saxon Glossography: Papers read at the International Conference held in the Koninklijke Academie voor Wetenschappen Letteren en Schone Kunsten van België, Brussels, 8 and 9 September 1986*. Ed. R. Derolez. Brussels: Paleis der Academiën, 1992. 77–96.
- Pierazzo, Elena. "Digital Genetic Editions: The Encoding of Time in Manuscript Transcription." *Text Editing, Print and the Digital World*. Ed. Marilyn Deegan and Kathryn Sutherland. Aldershot: Ashgate, 2009. 169–186.
- Pierazzo, Elena and Malte Rehbein. *Documents and Genetic Criticism TEI Style*. TEI Consortium, 2010. <<http://www.tei-c.org/SIG/Manuscripts/genetic.html>>.
- Prescott, Andrew. "'Their Present Miserable State of Cremation': The Restoration of the Cotton Library." *Sir Robert Cotton as Collector: Essays on an Early Stuart Courtier and his Legacy*. Ed. Christopher John Wright. London: British Library Publications, 1997. 391–454.
- Pulliam, Heather. *Word and Image in the Book of Kells*. Dublin: Four Courts Press, 2006.
- Renear, Allen, Elli Mylonas and David Durand. "Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies." *Research in Humanities Computing*. Eds. Nancy Ide and Susan Hockey. Oxford: Oxford University Press, 1996. <<http://www.stg.brown.edu/resources/stg/monographs/ohco.html>>.
- Renear, Allen H. "Text Encoding." *A Companion to Digital Humanities*. Eds. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell, 2004. 218–239. <<http://www.digitalhumanities.org/companion/>>.
- Robinson, Pamela R. "The 'Booklet': A Self-Contained Unit in Composite Manuscripts." *Codicologica*. Eds. Albert Gruys and J. Peter Gumbert. 3 vols. Leiden: Brill, 1980. 46–69.
- Rollason, David and Lynda Rollason, eds. *The Durham Liber Vitae ... The Complete Edition*. 3 vols. London: The British Library, 2007.
- Rumble, Alexander R. "The palaeography of the Domesday manuscripts." *Domesday book: a reassessment*. Ed. Peter H. Sawyer. London: E. Arnold, 1985. 28–49.

- Sardou, Victorien, Giuseppe Giacosa, Luigi Illica and Giacomo Puccini. *Tosca. Facsimile della copia di lavoro del libretto*. Ed. Gabriella Biagi Ravenni. 2 vols. Firenze: Centro Studi Giacomo Puccini – Leo Olschki Editore, 2009.
- Sauer, Hans. "The Transmission and Structure of Archbishop Wulfstan's 'Commonplace Book'". *Old English Prose: Basic Readings*. Ed. Paul E. Szarmach. New York: Garland, 2000. 339–93.
- Sawyer, Peter H., ed. *Domesday book: a reassessment*. London: E. Arnold, 1985.
- Schillingsburg, Peter L. *From Gutenberg to Google*. Cambridge: Cambridge University Press, 2006.
- Sisam, Kenneth. *Studies in the History of Old English Literature*. Oxford: Clarendon Press, 1953.
- Sperberg-McQueen, C. Michael "Text in the Electronic Age: Textual Study and Text Encoding with Examples From Medieval Text." *Literary and Linguistic Computing*, 6 (1991): 32–46.
- Stinson, Timothy. "Codicological Descriptions in the Digital Age." *Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Eds. Malte Rehbein, Patrick Sahle and Torsten Schaßan. Norderstedt: Books on Demand, 2009. 35–51.
- Stokes, Peter Anthony. "The Vision of Leofric: Manuscript, Text and Content." *Review of English Studies*, (forthcoming).
- Suntrup, Rudolf. "'Te igitur'-Initialen und Kanonbilder in mittelalterlichen Sakramentarhandschriften." *Text und Bild: Aspekte des Zusammenwirkens zweier Künste in Mittelalter und früher Neuzeit*. Eds. Christel Meier and Uwe Ruberg. Wiesbaden: Ludwig Reichert, 1980. 278–382.
- Sutherland, Kathryn. *Jane Austen's textual lives: from Aeschylus to Bollywood*. Oxford: Oxford University Press, 2005.
- Tardieu, Jean. *Poème à voir*. Paris: Gallimard, 1990.
- TEI P5: *Guidelines for Electronic Text Encoding and Interchange*. 1.7.0, 2010-07-07, TEI Consortium, 2009. <<http://www.tei-c.org/Guidelines/P5/>>
- Tinti, Francesca. "From Episcopal Conception to Monastic Compilation: Hemming's Cartulary in Context." *Early Medieval Europe*, 11 (2002): 233–61.
- Tinti, Francesca. "Si litterali memorie commendaretur: Memory and Cartularies in Eleventh-Century Worcester." *Early medieval studies in memory of Patrick Wormald*. Ed. Stephen Baxter et al. Farnham: Ashgate, 2009. 475–91.
- Treharne, Elaine M. "The Bishop's Book: Leofric's Homiliary and Eleventh-Century Exeter." *Early medieval studies in memory of Patrick Wormald*. Ed. Stephen Baxter et al. Farnham: Ashgate, 2009. 521–37.
- Whitman, Walt. "America to Old-World Bards (Poetry Manuscript)." *Walt Whitman Archive*. 1995–2010. Eds. Ed Folsom and Kenneth M. Price. Lincoln (NE): Center for Digital Research in the Humanities at the University of Nebraska–Lincoln, 1995–2010. <<http://www.whitmanarchive.org/>>.

Appendices

Kurzbiographien – Biographical Notes

Philippe Artières est chercheur en histoire au CNRS à l'Ecole des Hautes Etudes en Sciences Sociales à Paris. Il est membre de l'Equipe Anthropologie de l'écriture de l'Institut Interdisciplinaire d'Anthropologie du Contemporain. Il est l'auteur d'une série de travaux sur l'histoire contemporaine de l'écriture. Tentant la synthèse entre une histoire matérielle de l'écriture et une histoire sociale, il prépare une étude sur Thérèse de Lisieux, une vie écrite et une analyse de l'enseigne lumineuse à paraître tous les deux fin 2010.

Bernhard Assmann studierte an der Universität zu Köln Informationsverarbeitung, Mittlere und Neuere Geschichte und Historische Hilfswissenschaften. Danach betreute er das Digitalisierungsprojekt "Die Werke Friedrichs des Großen" an der Universitätsbibliothek Trier. Gegenwärtig ist er beim Hochschulbibliothekszenrum des Landes Nordrhein-Westfalen in Köln beschäftigt.

Jean François Bert est sociologue, spécialiste de l'histoire de la sociologie et de l'anthropologie française durant la 1er moitié du XXe siècle. Il est l'un des animateurs du Centre Michel Foucault et s'intéresse à la réception du philosophe dans les sciences humaines.

Pierre-Yves Buard is a PhD student at the University of Caen Basse-Normandie, member of the GREYC laboratory (CNRS UMR 6072) and engineer at the Caen University Press. His work focuses on connections between scholarly publishing and digitalization.

Toby Burrows has a PhD from the University of Western Australia and an M.A. from the University of London, both in medieval history. From 2005 to 2010 he was the Digital Services Director of the Australian Research Council's Network for Early European Research. He is now working in research data management at the University of Western Australia as well as managing the University Library's rare book and manuscript collections.

Yaacov Choueka, Professor Emeritus of Computer Science, got his PhD in Mathematics from the Hebrew University in Jerusalem in 1971, and was affiliated with the Department of Mathematics and then of Computer Science of Bar Ilan University in Ramat Gan till 2004. His research interests focused on Natural Language Processing and on the intelligent interface between computers and linguistic corpora in particular. He initiated and headed several very large scale projects in these areas, and was awarded Israel Prime Minister Prize for Computing in 1997. He is currently

heading the Genazim team for the computerization of the Genizah collection of Hebrew manuscripts, and leading its research team for the computerized analysis of digitized historical manuscripts.

Julia Craig-McFeely is a research Fellow in the Faculty of Music at the University of Oxford. She has been Project Manager of the DIAMM Project since 1998 and one of its co-directors since 2005. Her doctoral dissertation (in historical musicology) was a codicological and palaeographical study of lute tablatures in England, but since then her research interests have extended to much earlier repertoires and broadened into the construction and application of databases and datasets to aid humanities research. She is a leading expert in the use of high-end digital imaging equipment, and was a consultant to the pilot project to digitize the Dead Sea Scrolls.

Daniel Deckers ist wissenschaftlicher Mitarbeiter und Koordinator des Teuchos-Zentrums am Institut für Griechische und Lateinische Philologie der Universität Hamburg. Er beschäftigt sich seit 2002 mit der Wiedergewinnung gelöschter Texte in griechischen Palimpsesten und kooperiert seit 2005 mit dem Hamburger Synchrotronstrahlungslabor zur Röntgenfluoreszenzanalyse von Handschriften. Seine Forschungsschwerpunkte sind Antikes Schriftwesen, multispektrale Handschriftendigitalisierung, gräzistische Palimpsestforschung, digitale Klassische Philologie und Editorik.

Nachum Dershowitz is a professor of computer science at Tel Aviv University. In addition to his interest in the analysis of historical texts, he is engaged in research in natural language processing, the theory of rewriting, models of computation, program verification, and calendrical algorithms.

Markus Diem has finished his studies on “Computer Graphics and Digital Image Processing” at the Vienna University of Technology in 2010. He is currently engaged as a research assistant for the Document Information Retrieval project (DIR) at the Computer Vision Lab, Institute of Computer Aided Automation, Vienna University of Technology, Austria.

Carole Dornier is Professor (French literature, XVIIIth century) at the University of Caen Basse-Normandie. Author of works on Crebillon fils, Prévost, Duclos, Rousseau, Robert Challes and Montesquieu, she has published critical editions of Duclos and Montesquieu. She directs a digital edition of “Mes Pensées” de Montesquieu (Projet Montedite: CRHQ, UMR-CNRS 6583 et Presses Universitaires de Caen).

Paolo Eleuteri è Professore di Codicologia all'Università Ca' Foscari di Venezia. Interessi scientifici principali: scritture greche, catalogazione dei manoscritti, storia e ricezione dei testi.

Franz Fischer is Post-Doctoral Researcher for the St Patrick's Confessio Hypertext Stack Project at the Royal Irish Academy, Dublin. He studied Latin and History in Cologne and in Rome. For his PhD he created a digital edition of William of Auxerre's treatise on liturgy. He worked as a research associate at the Thomas Institute for Medieval Philosophy and at the Centre for Medieval Studies, University of Cologne, and held teaching positions for Latin and Digital Philology. A co-founder of the Institute for Documentology and Scholarly Editing, his research interest is focused on scholarly digital editions of medieval Latin texts.

Christiane Fritze hat Romanistik, Slawistik und Bibliotheks- und Informationswissenschaft studiert. Zurzeit ist sie wissenschaftliche Mitarbeiterin im europäischen Forschungsinfrastrukturprojekt DARIAH (Digital Research Infrastructure for the Arts and Humanities) an der Staats- und Universitätsbibliothek Göttingen. Zuvor war sie als wissenschaftliche Mitarbeiterin an der Berlin-Brandenburgischen Akademie der Wissenschaften mit verschiedensten Aspekten der Digitalisierung und Transformation geisteswissenschaftlicher Inhalte konfrontiert – so beim Digitalen Wörterbuch der deutschen Sprache, der Initiative Telota und als Koordinatorin beim Deutschen Textarchiv.

Melanie Gau has specialised in Computational Linguistics and Palaeoslavic Studies. Currently she is taking part in the Austrian Science Fund project "Critical Edition of the New Sinaitic Glagolitic Euchology (Sacramentary) Fragments with the Aid of Modern Technologies" and is writing her PhD-thesis on the "Psalterium Demetrii Sinaitici (Sin. slav. 3/N)".

Tanya German received an BSc degree in Biomedical Engineering from the Technion in 2005. Since 2009 she is an MSc Electrical Engineering student at the Tel Aviv University. Her areas of interest include signal and image processing.

Leif Glaser works as a PostDoc at the Petra III soft X-ray beamline P04 at DESY in Hamburg. His PhD covers circular magnetic dichroism measurements on magnetic CoPt clusters and nanoparticles. He measures historic writings with synchrotron radiation x-ray fluorescence in a cooperation with the Teuchos centre at the University of Hamburg since 2005. His most recent measurements using synchrotron based techniques are on a neolithic bronze axe.

Carmen Kämmerer studierte Germanistik, Romanistik und Anglistik (Staatsexamen und M.A.) an den Universitäten Mannheim und Heidelberg. Die Promotion erfolgte im Fachbereich ältere deutsche Sprachwissenschaft an der Universität Zürich. Von 2007 bis 2009 absolvierte sie das Bibliotheksreferendariat an der Herzog August Bibliothek in Wolfenbüttel und der Bayerischen Staatsbibliothek in München.

Bis Juli 2010 leitete sie die Handschriftenabteilung der Württembergischen Landesbibliothek in Stuttgart.

Robert Kummer studied Information Management at the Baden-Württemberg Cooperative State University Stuttgart and Humanities Computer Science at the University of Cologne. Since 2003 he has been involved in several research projects in the areas of database development, information integration, preservation and long-term access. Currently he is working on his PhD dissertation project dealing with semantic information integration of cultural heritage information systems.

Marilena Maniaci è docente di Storia del libro manoscritto presso l'Università degli studi di Cassino. Principali interessi di ricerca: materiali e tecniche di manifattura del libro medievale greco e latino (supporti; confezione dei fascicoli; preparazione della pagina e 'mise en page'; codici 'complessi'; caratteri materiali di specifiche tipologie librarie); terminologia del manoscritto.

Volker Märgner studierte Elektrotechnik an der Technischen Universität Braunschweig (TUBS). Das Studium schloss er 1974 mit dem Diplom und 1983 mit der Promotion zum Dr.-Ing. ab. Seit 1976 ist er am Institut für Nachrichtentechnik der TUBS tätig. Zurzeit hat er die Position eines Akademischen Direktors inne und arbeitet in Lehre und Forschung im Bereich der Digitalen Signalverarbeitung, insbesondere der Bildverarbeitung und Mustererkennung. Seine aktuellen wissenschaftlichen Interessen konzentrieren sich auf Erkennung handgeschriebener Wörter in historischen Dokumenten und Materialprüfung mittels Online-Thermographie.

Ulrike Mehringer ist Diplom-Bibliothekarin und arbeitet seit 1992 an der Universitätsbibliothek Tübingen, wo sie seit 2006 im Bereich "Historischer Lesesaal, Handschriften und Alte Drucke" beschäftigt ist.

Peter Meinlschmidt studierte Physik an den Universitäten in Oldenburg und Hamburg sowie an der Towson State University (USA). Seit 1996 ist er als Projektleiter am WKI auf dem Gebiet zerstörungsfreier Techniken tätig. Seine aktuellen wissenschaftlichen Interessen sind optische Messtechniken, Thermographie, Holografie, Spectral Imaging und Bildverarbeitung.

Heinz Miklas is professor at the Institute of Slavic Studies at the University of Vienna. He is mainly engaged in the comparative history of Slavic writing systems, Slavonic codicology and palaeography and the edition of Glagolitic and Cyrillic texts. He is head of the Austrian Science Foundation project "Critical Edition of the New Sinaitic Glagolitic Euchology (Sacramentary) Fragments with the Aid of Modern Technologies". He is a member of the Balkan-Commission/AAS (since 1995), Board of directors, then Supervisory board of the Austrian Institute for East and

South-East European Studies (1999-2006), Supervisory board of the Institute for the Danube Region and Central Europe (IDM, since 2007) and founder of the Wiener Archäographisches Forum.

Pádraig Ó Macháin is a Professor at the School of Celtic Studies, Dublin Institute for Advanced Studies. His research interests include the manuscript tradition of Gaelic Ireland, and the language and literature of late medieval and early modern Ireland. He is director of Irish Script on Screen.

Ezio Ornato est Directeur de recherche émérite au Centre national de la recherche scientifique, auquel il a appartenu de 1962 à 2001, et travaille à Villejuif dans le Laboratoire de médiévistique occidentale de Paris (LAMOP). Ayant reçu une formation dans le domaine de la philologie humaniste, il a tout d'abord publié l'œuvre de l'humaniste français Jean de Montreuil (1353-1418), puis il a travaillé sur l'histoire de la tradition manuscrite des discours de Cicéron. Depuis une quarantaine d'années, son intérêt s'est concentré sur l'histoire du livre — tant manuscrit qu'imprimé — ainsi que sur l'histoire du papier filigrané.

Elena Pierazzo has a PhD in Italian Philology: her specialism is Italian Renaissance texts and text encoding and she has published and presented papers at international conferences in Renaissance literature, digital critical editions, text encoding theory and Italian linguistics. She is currently an Associate Researcher at the Centre for Computing in the Humanities at King's College London, and lead analyst of a dozen research projects; she is also a teacher of XML-related technologies at both undergraduate and Masters level. She is the chair of the TEI Manuscripts SIG with Malte Rehbein and an elected member of the TEI Council since 2007. She is one of the members of the MS-SIG task force that proposed a new TEI module for documentary and genetic editing.

Liza Potikha is an MSc Computer Science student at the Tel Aviv University. Her areas of interest include image processing, feature extraction and clusterization. She also works in industry as a research engineer and specializes in algorithms for video cameras.

Stephen Quirke is curator at the Petrie Museum of Egyptian Archaeology and Professor of Egyptian Archaeology at the Institute of Archaeology, University College London (UCL). At the Petrie Museum he curated the digitisation and online illustrated catalogue of all 80,000 objects. He is a specialist in the ancient Egyptian cursive script ('hieratic'), with historical focus on the late Middle Kingdom (1850-1700 BC): his Cambridge PhD was on administrative documents of that period, and with Mark Collier from 2002-2006 he published its largest varied manuscript dataset, the UCL Lahun Papyri.

Malte Rehbein is lecturer in Digital Humanities and History and coordinator of the Centre for Digital Editions at the University of Würzburg. His main research is on methodology for encoding and processing complex texts, on textual variation, genetic editing and scholarly digital editions as well as manuscript studies. He has a doctorate in Medieval History from the University of Göttingen. He is a former Marie Curie research fellow of the National University of Ireland, Galway, co-chair of the TEI Manuscripts Special Interest Group, elected member of the executive board for the Digital Medievalist and editor-in-chief of the Digital Medievalist Journal.

Martin Rüesch studied history, philosophy and constitutional law at the University of Zurich. He is working on a PhD thesis on the influence of Pierre Bayle's *Dictionnaire historique et critique* with regard to changes in information culture of the late 17th and early 18th century. From 2005 to 2008 he was researching and teaching at the Department of History at the University of Heidelberg. Since 2009 he has been the head of the e-learning project "Ad fontes" at the University of Zurich.

Robert Sablatnig is an associate professor of computer vision heading the Computer Vision Lab (which was part of the Pattern Recognition and Image Processing Group), and is head of the Institute of Computer Aided Automation, engaged in research, project leading, and teaching. His research interests are 3D Computer Vision, Automatic Visual Inspection, Hierarchical Pattern Recognition, Video data analysis, Automated Document Analysis, Multispectral Imaging, Virtual- and Augmented Reality, and Applications in Industry and Cultural Heritage Preservation. He is Vice President of the Austrian Association for Pattern Recognition (AAPR/OAGM), the Austrian branch of IAPR.

Patrick Sahle is still a Lecturer in Humanities IT at the University of Cologne where he is now—and this is only the update to the information given in the previous volume on Codicology and Palaeography—also working for the Cologne Center for eHumanities (CCeH), a research and education center in the Faculty of the Humanities.

Samantha Saïdi est ingénieure d'études à l'Ecole normale supérieure de Lyon où elle travaille pour le laboratoire Triangle - UMR 5206 depuis 2005. Elle participe à ce titre aux projets d'éditions critiques ou de corpus du laboratoire, tels l'"Echo de la Fabrique et la petite presse ouvrière lyonnaise des années 1831-1835", et "Tchitchérine, le libéralisme en Russie : constitution d'un corpus numérique en cyrillique". Elle co-anime avec des collègues de l'Ecole normale supérieure et de l'Institut des sciences de l'homme, un dispositif de partage, d'accumulation et de

diffusion des technologies et des méthodologies qui émergent dans le champ des humanités numériques : Mutualisation pour les éditions critiques et les corpus (MutEC).

Silke Schöttle ist Diplom-Archivarin (FH). Sie war 1998-2001 Kreisarchivarin des Alb-Donau-Kreises in Ulm und hat 2001-2007 das Studium der Geschichte und Romanischen Philologie in Tübingen, Barcelona und Salamanca absolviert. 2007-2010 war sie wissenschaftliche Mitarbeiterin der Universitätsbibliothek Tübingen zur Erschließung der deutschen Handschriften (16.-20. Jh.). Seit April 2010 ist sie als wissenschaftliche Mitarbeiterin des Landesarchivs Baden-Württemberg tätig.

Isabelle Schürch studied history, English literature and psychology at the University of Berne. In 2008 she worked as a scientific researcher at the State Archive of Argovia. From fall 2008 to spring 2010 she was employed as an assistant at the Department of History (Chair of Prof. Dr. Simon Teuscher, Medieval History) which enabled combining teaching and archival studies. Since 2008 she has been working on a PhD thesis on "Power and Information. Missives as Media of Lordship in the Late Middle Ages" which now forms part of the interdisciplinary NCCR "Mediality" at the University of Zurich.

Roni Shweka has a BSc degree in computer sciences (1995) and a PhD degree in Talmud from the Hebrew University of Jerusalem (2009). His main research area is the Rabbinical literature at the Geonic period, from the 8th century to the 11th century. His research interests however are involved with many other aspects of the huge Cairo Genizah collection. From 2007 he is working as a Genizah expert in Genazim, the computerization unit of The Friedberg Genizah Project. Shweka is also an assistant-editor of the Catalog of Talmudic Manuscripts to be soon published.

Ken Sochats is the Carnegie Science Center 2002 award winner for IT excellence. He holds advanced degrees in Electrical Engineering and Business Administration. He is currently the Director of the Visual Information Systems Center at the University of Pittsburgh. He has over thirty years of experience in the Education, Computer and Telecommunications industries. His work at Westinghouse Electric Corporation resulted in several inventions and patents. He served as Vice President of Information Systems at BroadStreet Communications Corporation. He was the manager of the Link To Learn Project out of The Governor's Office of Information Technology.

Timothy Stinson is an assistant professor of English at North Carolina State University and co-editor of "The Siege of Jerusalem Electronic Archive", which is forthcoming from the Society for Early English and Norse Electronic Texts. He has held research grants or fellowships from the National Endowment for the Humanities, the

Andrew W. Mellon Foundation, the Council on Library and Information Resources, and the Bibliographical Society of America.

Peter A. Stokes is Research Fellow at the Centre for Computing in Humanities, King's College London, and Principal Investigator of the new "Digital Resource and Database of Palaeography, Manuscripts and Diplomatic", funded by a European Research Council Starting Grant (FP7). As well as palaeographical method and the application of computing to manuscript studies, his other primary interests include the vernacular English scripts of the late-tenth through twelfth centuries. He has also published on computing in lexicography, Anglo-Saxon charters and bounds, and early-modern book collectors, and has developed software for digital humanities.

Alison Stones is Professor of History of Art and Architecture at the University of Pittsburgh and teaches medieval art and architecture. Her research is on illuminated manuscripts of France, Italy and Spain from the 12th to 15th centuries.

Dominique Stutzmann est docteur en histoire, chargé de recherche au Centre National de la Recherche Scientifique (UPR 841 – Institut de Recherche et d'Histoire des Textes, section de paléographie) et chargé de conférences en paléographie médiévale à l'École Pratique des Hautes Études. Il a étudié les lettres classiques, l'allemand et l'histoire à la Sorbonne (universités Paris 1 et Paris 4) et obtenu le diplôme d'archiviste paléographe de l'École nationale des Chartes. Il a été conservateur à la Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (département des Manuscrits) et à la Bibliothèque nationale de France (département de l'Information bibliographique et numérique).

Melissa Terras is the Reader in Electronic Communication in the Department of Information Studies, University College London. With a background in Classical Art History and English Literature, and Computing Science, her doctorate (University of Oxford) examined how to use advanced information engineering technologies to interpret and read the Vindolanda texts. She is a general editor of DHQ, and the Secretary of the Association of Literary and Linguistic Computing. Her research focuses on the use of computational techniques to enable research in the arts and humanities that would otherwise be impossible.

Armand Tif ist Kunsthistoriker und seit 2010 an der Österreichischen Akademie der Wissenschaften tätig. Nach dem Studium absolvierte er den zweijährigen postgradualen Kurs »Curriculum eCompetence« und war 2006-2010 Lehrbeauftragter und wissenschaftlicher Mitarbeiter der Universität Wien. Im Rahmen des Projekts »Entwicklung digitaler Lehrmodule für die Studieneingangsphase (Kunstgeschichte)« wirkte er 2006–2007 an der Erstellung von eLearning-Systemen

für Überblicksvorlesungen zu den Epochen der Kunst mit. 2007–2010 war er im Katalogisierungsprojekt der illuminierten Handschriften und Inkunabeln der Österreichischen Nationalbibliothek am Otto Pächt-Archiv beschäftigt.

Georg Vogeler ist Lehrbeauftragter und Habilitand an der Ludwig-Maximilians-Universität München (LMU). Er hat 2002 über spätmittelalterliche Steuerbücher deutscher Territorien promoviert, war Fedor-Lynen-Stipendiat der Alexander-von-Humboldt-Stiftung an der Università del Salento in Lecce, bis 2010 wissenschaftlicher Assistent am Lehrstuhl für Geschichtliche Hilfswissenschaften der LMU. Er ist Mitglied im Moderamen der Association Paléographique Internationale – Culture, Écriture, Société (APICES) und technischer Direktor des International Center for Archival Research (ICARus). Er arbeitet über spätmittelalterliches Verwaltungsschriftgut, das Urkundenwesen Kaiser Friedrichs II. und den Einsatz digitaler Methoden in der Diplomatik.

Lior Wolf is a faculty member at the Computer Science Department at Tel-Aviv University. Previously, he was a post-doctoral associate in Prof. Poggio's lab at MIT. He graduated from the Hebrew University, Jerusalem, where he worked under the supervision of Prof. Shashua. Lior Wolf was awarded the 2008 Sackler Career Development Chair, the Colton Excellence Fellowship for new faculty (2006–2008), the Max Shlumiuk award for 2004, and the Rothchild fellowship for 2004. His joint work with Prof. Shashua in ECCV 2000 received the best paper award, and their work in ICCV 2001 received the Marr prize honorable mention. He was also awarded the best paper award at the post ICCV 2009 workshop on eHeritage.

KPDZ 1 – CPDA 1

Kodikologie und Paläographie im Digitalen Zeitalter / Codicology and Palaeography in the Digital Age, hg. v. Malte Rehbein, Patrick Sahle und Torsten Schaßan unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009.

ISBN 978-3-8370-9842-6

Online: <<http://kups.ub.uni-koeln.de/volltexte/2009/2939/>>

Der gedruckte Band kann zum Preis von € 49,- über den Buchhandel, über amazon.de und über die Webseite des Verlages <http://www.bod.de/index.php?id=1132&objk_id=217805> bezogen werden.

You can order the printed version at the price of € 49,- from your local bookstore, via amazon.de or via the website of the publishing house: <http://www.bod.de/index.php?id=1132&objk_id=217805>.

Beiträge – Contributions

Georg Vogeler: Einleitung. Der Computer und die Handschriften

Francesco Bernardi, Paolo Eleuteri, Barbara Vanin: La catalogazione in rete dei manoscritti delle biblioteche venete: *Nuova Biblioteca Manoscritta*

Antonio Cartelli, Andrea Daltari, Paola Errani, Marco Palma, Paolo Zanfini: Il catalogo aperto dei manoscritti Malatestiani

Christian Speer: Die Sammlung Georg Rörers (1492–1557). Ein interdisziplinäres und multimediales Erschließungsprojekt an der Thüringer Universitäts- und Landesbibliothek Jena

Timothy Stinson: Codicological Descriptions in the Digital Age

Pamela Kalning, Karin Zimmermann: Die Digitalisierung der deutschsprachigen Handschriften der Bibliotheca Palatina in der Universitätsbibliothek Heidelberg

Zdeněk Uhlíř, Adolf Knoll: Manuscriptorium Digital Library and ENRICH Project: Means for Dealing with Digital Codicology and Palaeography

Daniel Deckers, Lutz Koch, Cristina Vertan: Representation and Encoding of Heterogeneous Data in a Web Based Research Environment for Manuscript and Textual Studies

Christina Wolf: Aufbau eines Informationssystems für Wasserzeichen in den DFG-Handschriftenzentren

Silke Kamp: Handschriften lesen lernen im digitalen Zeitalter

Antonio Cartelli, Marco Palma: Digistylus — An Online Information System for Palaeography Teaching and Research

Bernard J. Muir: Innovations in Analyzing Manuscript Images and Using them in Digital Scholarly Publications

Hugh A. Cayless: Linking Text and Image with SVG

Patrick Shiel, Malte Rehbein, John Keating: The Ghost in the Manuscript: Hyperspectral Text Recovery and Segmentation

Daniele Fusi: Aspects of Application of Neural Recognition to Digital Editions

Gilbert Tomasi, Roland Tomasi : Approche informatique du document manuscrit

Arianna Ciula: The Palaeographical Method Under the Light of a Digital Approach

Mark Stansbury: The Computer and the Classification of Script

Maria Gurrado: «Graphoskop», uno strumento informatico per l'analisi paleografica quantitativa

Wernfried Hofmeister, Andrea Hofmeister-Winter, Georg Thallinger: Forschung am Rande des paläographischen Zweifels: Die EDV-basierte Erfassung individueller Schriftzüge im Projekt DAMaLS

Mark Aussems, Axel Brink: Digital Palaeography

Peter A. Stokes: Computer-Aided Palaeography, Present and Future